DOCUMENT RESUME

ED 458 220 TM 033 373

AUTHOR	Bourque, Mary Lyn, Ed.
TITLE	Proceedings of Achievement Levels Workshop (Boulder, Colorado, August 20-22, 1997).
INSTITUTION	National Assessment Governing Board, Washington, DC.
PUB DATE	1998-00-00
NOTE	118p.
AVAILABLE FROM	National Assessment Governing Board, 800 North Capitol Street, NW, Suite 825, Washington, DC 20002-4233. For full text: http://www.nagb.org.
PUB TYPE	Collected Works - Proceedings (021)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Academic Achievement; Elementary Secondary Education; Item Response Theory; National Competency Tests; *Standards; *Test Construction; *Validity
IDENTIFIERS	National Assessment Governing Board; *National Assessment of Educational Progress; Standard Setting

ABSTRACT

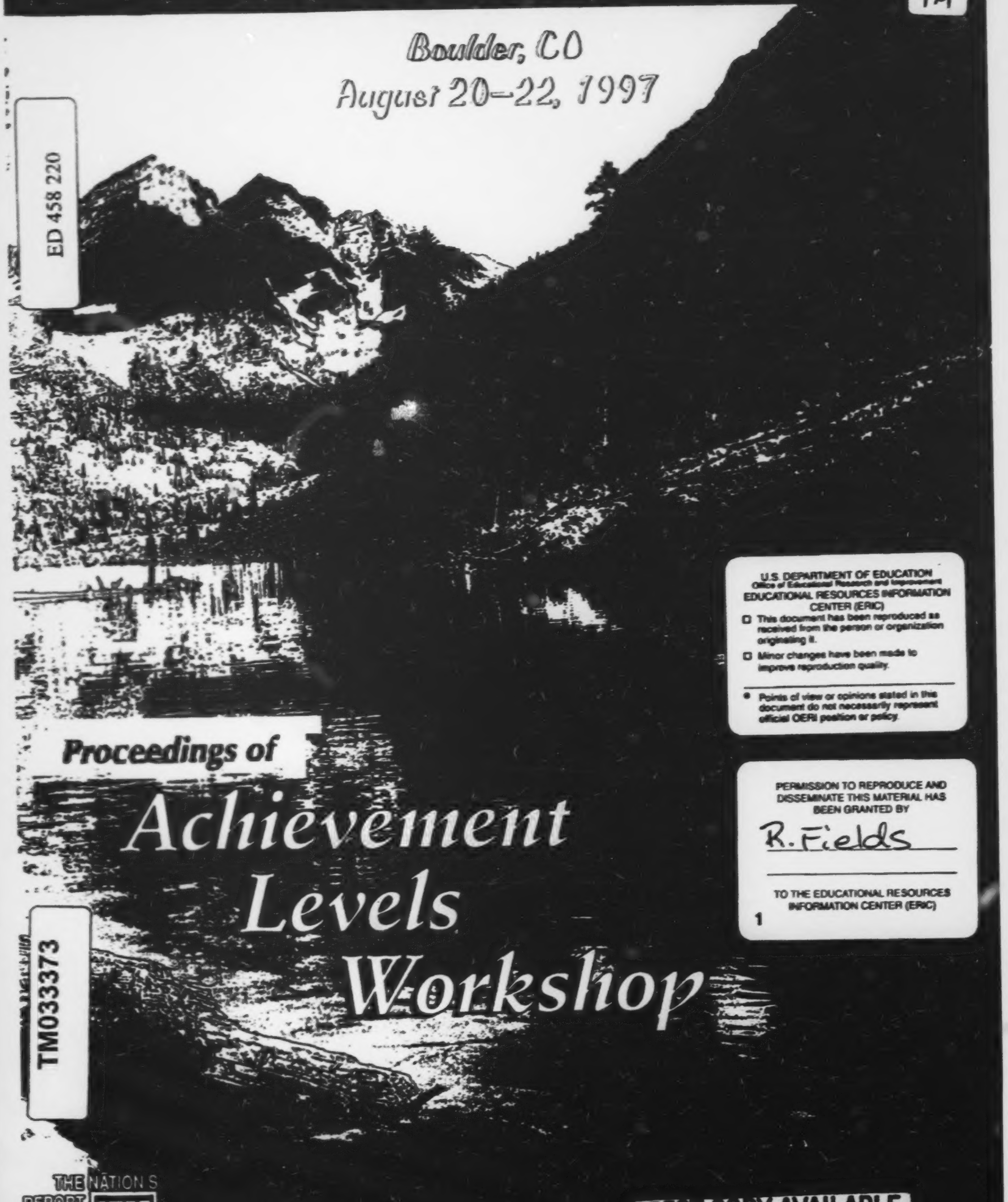
The papers presented at this symposium explore each component of the National Assessment of Educational Progress (NAEP), from the development of the assessment frameworks through the reporting of assessment results. The papers are: (1) "NAEP Frameworks and Achievement Levels" (Robert A Forsyth); (2) "Assembly of Test Forms for Use in Large-Scale Educational Assessments" (Wim J. van der Linden); (3) "A Brief Introduction to Item Response Theory for Items Scored in More Than Two Categories" (David Thissen, Kathleen Billeaud, Lori McLeod, and Lauren Nelson); (4) "Some Ideas about Item Response Theory Applied to Combinations of Multiple-Choice and Open-Ended Items: Scale Scores for Patterns of Summed Scores" (Kathleen Billeaud, Kimberly Swygert, Lauren Nelson, and David Thissen); (5) "Enhancing the Validity of NAEP Assessment Level Score Reporting" (Ronald K. Hambleton); (6) "1998 Civics and Writing Level-Setting Methodologies" (ACT, Inc.); (7) "The Criticality of Consequences in Standard Setting: Six Lessons Learned the Hard Way by a Standard-Setting Abettor" (W. James Popham); and (8) "Acknowledgments and Appendices." Two appendixes list participants and present the conference agenda. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

Boulder, CO
August 20–22, 1997

ED 458 220



Proceedings of

Achievement Levels Workshop

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
R. Fields

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

TM033373

THE NATION'S REPORT CARD
naep

National Assessment Governing Board
U.S. Department of Education

BEST COPY AVAILABLE

What Is The Nation's Report Card?

The Nation's Report Card, the National Assessment of Educational Progress (NAEP), is the only nationally representative and continuing assessment of what America's students know and can do in various subjects. Since 1969, assessments have been conducted periodically in reading, mathematics, science, writing, history/geography, and other fields. By making objective information on student performance available to policymakers at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement is collected under this program. NAEP guarantees the privacy of individual students and their families.

NAEP is a congressionally mandated project of the National Center for Education Statistics, the U.S. Department of Education. The Commissioner of Education Statistics is responsible, by law, for carrying out the NAEP project through competitive awards to qualified organizations. NAEP reports directly to the Commissioner, who is also responsible for providing continuing reviews, including validation studies and solicitation of public comment, on NAEP's conduct and usefulness.

The National Assessment Governing Board (NAGB) was established under section 412 of the National Education Statistics Act of 1994 (Title IV of the Improving America's Schools Act of 1994, P.L. 103–382). The Board was established to formulate policy guidelines for NAEP. The Board is responsible for selecting subject areas to be assessed, developing assessment objectives, identifying appropriate achievement goals for each grade level and subject tested, and establishing standards and procedures for interstate and national comparisons.

The National Assessment Governing Board

Mark D. Musick, Chair (1999)
President
Southern Regional Education Board
Atlanta, Georgia

Mary R. Blanton, Vice Chair (1998)
Attorney
Salisbury, North Carolina

Patsy Cavazos (1998)
Principal
W.G. Love Accelerated School
Houston, Texas

Catherine A. Davidson (1998)
Secondary Education Director
Central Kitsap School District
Silverdale, Washington

Edward Donley (1997)
Former Chairman
Air Products & Chemicals, Inc.
Allentown, Pennsylvania

James E. Ellingson (1998)
Fourth-Grade Classroom Teacher (Retired)
Probstfield Elementary School
Moorhead, Minnesota

Honorable John M. Engler,
(Member Designate)
Governor of Michigan
Lansing, Michigan

Thomas H. Fisher (1999)
Director
Student Assessment Services
Florida Department of Education
Tallahassee, Florida

Michael J. Guerra (2000)
Executive Director
National Catholic Education Association
Secondary School Department
Washington, D.C.

Edward H. Haertel (1999)
Professor
School of Education
Stanford University
Stanford, California

Lynn Marmer
President
Cincinnati Board of Education
Cincinnati, Ohio

William J. Moloney (1999)
Commissioner of Education
State of Colorado
Denver, Colorado

Honorable Annette Morgan (1998)
Former Member
Missouri House of Representatives
Jefferson City, Missouri

Mitsugi Nakashima (2000)
First Vice Chair
Hawaii State Board of Education
Honolulu, Hawaii

Michael T. Nettles (1999)
Professor of Education and Public Policy
University of Michigan
Ann Arbor, Michigan
and
Director
Frederick D. Patterson
Research Institute
United Negro College Fund

Honorable Norma Paulus (1999)
Superintendent of Public Instruction
Oregon State Department of Education
Salem, Oregon

Jo Ann Pottorff (2000)
Member
Kansas House of Representatives
Wichita, Kansas

Honorable William T. Randall,
Chair (1998)
Former Commissioner of Education
State of Colorado
Denver, Colorado

Diane Ravitch (2000)
Senior Research Scholar
New York University
New York, New York

Honorable Roy Romer (1998)
Governor of Colorado
Denver, Colorado

Fannie L. Simmons (1998)
Math Coordinator
District 5, Lexington/Richland County
Ballentine, South Carolina

Adam Urbanski (1998)
President
Rochester Teachers Association
Rochester, New York

Deborah Voltz (1999)
Assistant Professor
Department of Special Education
University of Louisville
Louisville, Kentucky

Marilyn A. Whirry (1999)
Twelfth-Grade English Teacher
Mira Costa High School
Manhattan Beach, California

Dennie Palmer Wolf (2000)
Senior Research Associate
Harvard Graduate School of Education
Cambridge, Massachusetts

C. Kent McGuire (Ex-Officio)
Acting Assistant Secretary of Education
Office of Educational Research
and Improvement
U.S. Department of Education
Washington, D.C.

Boulder, Colorado
August 20–22, 1997

Proceedings of
Achievement Levels Workshop

Edited by

Mary Lyn Bourque
National Assessment Governing Board

THE NATION'S REPORT CARD
naep
National Assessment Governing Board
U.S. Department of Education

National Assessment Governing Board

Mark Musick
Chair

Mary R. Blanton
Vice Chair

Roy Truby
Executive Director

Mary Lyn Bourque
Assistant Director for Psychometrics
and Committee Staff

September 1998

Suggested Citation

Bourque, M.L. (Ed.)
Proceedings of Achievement Levels Workshop
Washington, DC: National Assessment Governing Board, 1998.

For more information
contact:

National Assessment Governing Board
202–357–6938

For ordering information on this report, write:

National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002–4233

This report is also available on the World Wide Web.
http://www.nagb.org

Table of Contents

Introduction v

Section 1 1
NAEP Frameworks and Achievement Levels
Robert A. Forsyth, University of Iowa

Section 2 27
Assembly of Test Forms for Use in Large-Scale Educational Assessments
Wim J. van der Linden, University of Twente, The Netherlands

Section 3 45
A Brief Introduction to Item Response Theory for Items Scored in More Than Two Categories
David Thissen, Kathleen Billeaud, Lori McLeod, Lauren Nelson,
University of North Carolina at Chapel Hill

Section 4 63
Some Ideas about Item Response Theory Applied to Combinations of Multiple-Choice and Open-Ended Items: Scale Scores for Patterns of Summed Scores
Kathleen Billeaud, Kimberly Swygert, Lauren Nelson, David Thissen,
University of North Carolina at Chapel Hill

Section 5 77
Enhancing the Validity of NAEP Achievement Level Score Reporting
Ronald K. Hambleton, University of Massachusetts, Amherst

Section 6 99
1998 Civics and Writing Level-Setting Methodologies
ACT, Inc., Iowa City, IA

Section 7 107
The Criticality of Consequences in Standard Setting: Six Lessons Learned the Hard Way by a Standard-Setting Abettor
W. James Popham, University of California, Los Angeles

Section 8 113
Acknowledgements and Appendices

Introduction

The National Assessment Governing Board's policy on student performance standards states that the achievement levels should influence all aspects of the National Assessment of Educational Progress (NAEP) assessments, from the development of the assessment frameworks through the reporting of assessment results. The purpose of these proceedings is to explore each component of the assessment, from drawing board to Boardroom, in order to understand more fully the relationships between the performance standards and these assessment components, and the mutual impacts they have on each other.

In Section 1 Robert Forsyth from the University of Iowa examines the relationship between the assessment frameworks and the levels. He discusses the general purposes of the frameworks and the preliminary achievement level descriptions, those statements of content that students should know and be able to do at each level. His paper also examines selected characteristics of the frameworks, e.g., item formats, breadth of coverage, and the complexity of the cognitive dimensions, and how such characteristics influence both the preliminary and final achievement level descriptions.

Wim van der Linden from the University of Twente in the Netherlands describes two test assembly procedures used in large-scale assessments in Section 2. One procedure assigns items to forms in units called blocks as in NAEP, while the second method assigns unique items from the item pool to test forms. The author discusses the relationship between the characteristics of the assessment, such as size of the item pool and the number of desired forms, and the recommended methodology for test assembly.

In Sections 3 and 4 David Thissen and his colleagues at the University of North Carolina at Chapel Hill provide an introduction to item response theory for scoring assessments such as NAEP which have both multiple choice items as well as multiple-scored items. The relationship between the score scales and the standard setting methodologies is important since setting performance standards is impacted by data coming from different item types. Most NAEP assessments use mixed item formats, except the writing assessment, which employs only extended constructed responses to prompts. How these assessments are scored and scaled can have significant consequences for the standard-setting results.

In Section 5 Ronald Hambleton from the University of Massachusetts at Amherst explores the issue of score reporting in NAEP, and presents the results of a small-scale study of the understandability of NAEP score reports. Hambleton also provides some guidance on how to improve NAEP score reports and comments on the usefulness of market-basket reporting for NAEP, an index similar to the Consumer Price Index.

Section 6 provides the reader with a look at the proposed methodologies for developing the student performance standards on the 1998 NAEP civics and writing assessments. This section outlines the key features of the 1998 proposal and the overarching principles for developing the levels.

Finally, in Section 7, W. James Popham, professor emeritus from the University of California at Los Angeles, provides a commentary on the criticality of consequences in standard-setting. In his inimitable style, Popham offers six lessons learned the hard way by a standard-setting abettor. Popham calls on his many years of experience in implementing standard setting initiatives for more than three dozen high-stakes tests for students, teachers, and administrators, and shares his wisdom on the subject.

The National Assessment Governing Board hopes that the issues raised by the authors, and the solutions proposed, will extend the national conversation on standard setting, and will benefit not only the National Assessment, but all those individuals and agencies whose responsibilities includes setting performance standards in various academic subjects.

NAEP Frameworks and Achievement Levels

Robert A. Forsyth University of Iowa

August 1997

NAEP Frameworks and Achievement Levels

The National Assessment of Educational Progress (NAEP) is often identified as the nation's most comprehensive and reliable indicator of student achievement. In a 1993 report, the National Academy of Education (NAE) noted that NAEP is "an unparalleled source of information about U.S. students' academic achievement in many important subject areas" (National Academy of Education, 1993b, p. xix).

Because NAEP uses a careful sampling design, employs stringent security measures, has a high participation rate (for grades 4 and 8), collects considerable collateral information, assesses important content domains, and provides trend data, it has become one of the key sources of information not only about student achievement but also about other aspects of education in the United States. During the past 30 years, NAEP has earned its reputation as "The Nation's Report Card." The development of assessments that have gained such acceptance by both educators and the public has not been a casual undertaking. Listed below is a simplified outline of the complex assessment process that NAEP follows to provide achievement information:

1. Develop content framework and assessment/exercise specifications.
2. Construct an item pool to fit the specifications.
3. Gather data related to item functioning.
4. Select items for the assessment.
5. Administer the assessment.
6. Analyze the assessment data.
7. Report assessment information.

Each component represents an important aspect of the overall process.[1] The components are interrelated, but represent markedly different types of activities. If any component is not well designed and implemented, the usefulness of the information provided will be questioned.

In 1989, the National Assessment Governing Board (NAGB) decided that the primary reporting mechanism for subsequent assessments should be a standards-based system. Specifically, three levels of achievement (Basic, Proficient, and Advanced) were to be used as the basis for reporting NAEP results. Inherent in such a reporting system is the need for descriptions of what students performing at these levels should be able to accomplish and a process to translate these descriptions into performance levels on the NAEP scale. Recent assessments, therefore, have incorporated procedures for reporting results by achievement levels.

...NAGB decided that the primary reporting mechanism ... should be a standards-based system.

This paper considers the relationship between the achievement levels and the content frameworks and assessment/exercise specifications of the NAEP process. The paper is divided into three major sections. The first section provides an overview of the purposes of the assessment frameworks and specifications and notes some general evaluations of both. The recent introduction of preliminary achievement levels into the overall assessment process is then considered. The final section identifies several specific characteristics of the framework specifications and provides a discussion of the impact these characteristics might have on the achievement levels. A concluding statement ends the paper.

---

[1] This general process is similar to that used in the development of any large-scale achievement test.

General Purposes of Frameworks

As noted above, the first activity in the NAEP assessment process is the development of the framework and specifications. Actually, two publications usually result from this activity. One of these is the **framework** for the assessment [e.g., *Geography Framework for the 1994 National Assessment of Educational Progress* (National Assessment Governing Board, 1994b)] and the other is the **specifications** for the assessment [e.g., *Geography Assessment and Exercise Specifications for the 1994 National Assessment of Educational Progress* (National Assessment Governing Board, 1994a)].[2] Typically, the specifications document represents an elaboration of the information in the framework document. For most assessment areas, this elaboration provides a more comprehensive definition of the content domain than is given in the framework document.[3]

"The framework represents a comprehensive overview of the most essential outcomes . . ."

This paper uses the term "framework" to refer to the information in both documents unless otherwise indicated.

The following passage, taken from the 1994 Geography Framework, describes the general purposes of frameworks:

> The framework represents a comprehensive overview of the most essential outcomes of students' geography education at the prescribed grade levels as determined by the consensus committees and by the testimony of numerous witnesses at three public hearings. Designed to guide the development of assessment instruments, the framework cannot encompass everything that is taught in geography in all of the nation's classrooms, much less everything that should be taught. Nevertheless, this broad and innovative framework attempts to capture the range of geography content and thinking skills that students should possess as they progress through school. The framework's content embraces the complex problems of modern life that students will inevitably encounter both inside and outside their classrooms. It should be viewed, therefore, both as a guide for assessment and a potential tool for crafting a relevant and contemporary geography curriculum as it reflects the discipline's involvement in the complexities of contemporary issues. (National Assessment Governing Board, 1994b, pp. 2–3)

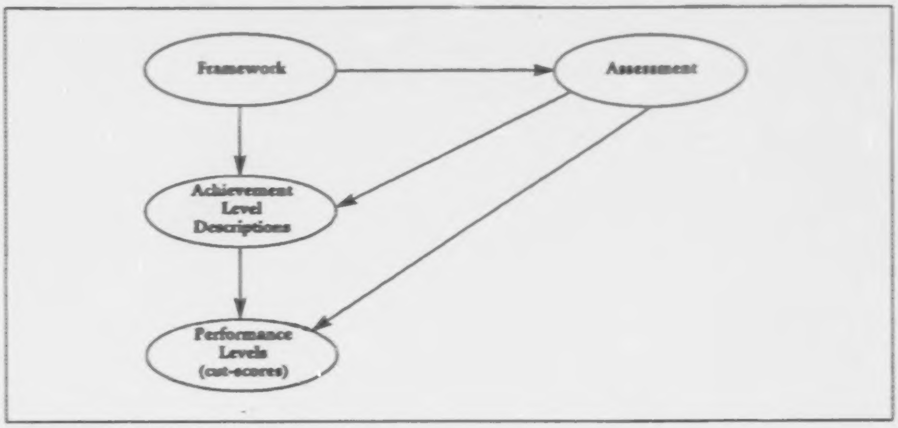
As indicated in this statement, frameworks provide detailed guidance for the construction of the exercise (item) pool. The types of stimulus material to be

---

[2] A single publication was developed for the 1998 Writing Assessment (National Assessment Governing Board, 1998b).

[3] For example, 58 (66%) of the 88 pages that make up the *Geography Assessment and Exercise Specifications* (National Assessment Governing Board, 1994a) are dedicated to detailed content specifications.

Figure 1.1. A simple model of the relationships among framework, achievement levels, and assessment


Framework
Assessment
Achievement Level Descriptions
Performance Levels (cut-scores)

used, the percentage of items in each unique combination of categories from the content and cognitive dimensions, the scoring criteria for constructed response exercises, and the percentage of testing time to be devoted to specific item formats are among the types of information provided in such frameworks. The statement does not, however, recognize another NAEP activity that depends greatly on the frameworks: the development of achievement levels. As the NAE notes:

> The framework must serve as the pivotal link between the assessment and the achievement levels—both as they are described narratively, and as they are operationalized into item-level judgments and ultimately cut-scores. (National Academy of Education, 1993a, p. 47)

Figure 1.1 shows the relationships among the frameworks, the assessment, and the achievement levels.[4]

In general, NAEP frameworks have been highly regarded by educators. In an evaluation of the 1992 Trial State Assessment, NAE concluded that the mathematics and reading frameworks were acceptable as a starting point for the assessment process. With respect to the mathematics framework, NAE states:

> The 1990 NAEP Mathematics Framework and the assessment design principles on which it was built were essentially sound. Furthermore, the framework represented a reasonable compromise between current instructional practices and the standards being put forth by the professional mathematics community at that time. (National Academy of Education, 1993b, p. 51)[5]

Similarly, with respect to the 1992 NAEP Reading Framework, NAE concludes:

---

[4] Figure 1.1 is similar to a figure published in National Academy of Education (1993a, p. 48). However, the NAE figure shows a reciprocal relationship between cut-scores and the assessment. This relationship did not seem reasonable, at least for initial assessments in an area. The NAE figure also does not indicate that the assessment has an impact on the achievement level descriptions.

[5] The 1992 Mathematics Assessment was also based on the 1990 framework.

> The reading consensus project produced a framework that represents a substantial advance over previous NAEP Reading Frameworks and is reasonably responsive to most of the current theories and practices in reading. (National Academy of Education, 1993b, p. 54)[6]

More recently, Mullis (1995, p. 3) observed that "the NAEP assessment frameworks are extremely well done and widely recognized for their breadth and depth of coverage." Likewise, Sireci (undated, p. 8) considered the consensus-building process used to develop the frameworks one of NAEP's great strengths.[7] Sireci also contended that "the content and cognitive domains are articulated clearly and are widely accepted by teachers, curriculum specialists, policymakers and other educational practitioners."

... the frameworks are ... a solid foundation ... (for) achievement levels descriptions.

If, in fact, the frameworks are as comprehensive and useful as indicated by the above statements, it seems reasonable to conclude that a solid foundation is in place to develop both the exercises and the achievement level descriptions. Of course, even the most critically acclaimed frameworks cannot guarantee that either the initial pool or the final set of exercises will adequately reflect the demands of the specifications. As Sireci notes: "The specifications of impressive content and cognitive frameworks is moot if the assessments do not adequately measure these frameworks" (p. 8).[8] Likewise, a similar statement could be made about the development of achievement level descriptions:

The availability of an excellent framework does not guarantee that useful achievement level descriptions will be developed.[9]

In recent NAEP assessments, preliminary achievement level descriptions have been included as part of the frameworks.[10] The characteristics and purposes of these preliminary descriptions are discussed in the next section.

Preliminary Achievement Level Descriptions

The 1994 U.S. History and Geography Assessments were the first to include Preliminary Achievement Level Descriptions (PALDs) as part of their frameworks (National Assessment Governing Board, 1994b, 1994e). Subsequently, PALDs also were incorporated into the frameworks for the 1996 Science Assessment (National

---

[6] The 1994 and 1998 Reading Frameworks are identical to the 1992 framework. A subsequent evaluation of the 1994 framework by NAE yielded a similar conclusion: "The expert advisors reaffirmed that the framework's general model of reading for meaning was consistent with current research practice and worked well as the basis for assessment" (National Academy of Education, 1996, p. 16).

[7] The Sireci paper was commissioned by the National Academy of Sciences. It seems clear that the paper was submitted in either 1996 or 1997.

[8] Later in his paper (pp. 57–59), Sireci suggests several content validity studies that should be done to investigate the congruence between the framework and the actual assessment. Linn and Dunbar (1992) have also suggested that more work needs to be done on content validity issues. When contrasting the efforts put forth to develop the frameworks to the efforts put forth to evaluate the items, they conclude: "We might be well served by focusing more of the attention of subject matter experts and educators on the individual exercises that make up an assessment" (p. 182).

[9] An example of this problem is identified in the NAE evaluation of the 1992 Reading Assessment (National Academy of Education, 1993a). As indicated previously, NAE considered the frameworks for this assessment to be adequate. However, the achievement levels were considered inadequate. NAE noted that "participants' lack of familiarity with the Reading Framework affected what they were able to hold in their mind when making item judgments and most certainly explains why the achievement level descriptions developed at the initial meeting had to be revised subsequently to be brought in line with the framework" (pp. 49–50).

[10] It should be noted that each framework since 1989 has also included what are usually referred to as "policy" or "generic" achievement level definitions that serve as the starting point for the subsequent achievement level definitions.

Assessment Governing Board, 1996b)[11] and the 1998 Civics and Writing Assessments (National Assessment Governing Board, 1998a, 1998b).

The lack of congruence between the achievement level descriptions and the exercise pool was a major criticism of earlier assessments and was probably a factor related to the introduction of PALDs as part of the frameworks. (See, for example, National Academy of Education, 1993b; Shepard, 1995.) In addition, given the pivotal role of the framework in developing achievement levels, it seems reasonable for content experts who develop the frameworks to be involved in the achievement level-setting process to some extent.

The importance placed on these PALDs with respect to the exercise-development component of the assessment process is illustrated by the statements below from the U.S. History and Geography Frameworks:[12]

> U.S. History:
> Exercises must be developed in such a way as to ensure that the item pool is congruent with the framework and corresponds to the achievement level descriptions. (National Assessment Governing Board, 1992, p. 12)
>
> Geography:
> The item pool should be developed in such a way as to ensure that the content described in the achievement level definitions … is reflected at each grade level. (National Assessment Governing Board, 1994a, p. 14)

These statements quite explicitly define the purpose of PALDs: to ensure an adequate exercise pool. However, it should be noted that these PALDs implicitly serve a second purpose: to guide the panels of educators and noneducators who will set the Final Achievement Level Descriptions (FALDs). These panels are convened after the assessment has been administered, and members of the original framework committees who established PALDs do not serve on them.[13] Under these circumstances, PALDs and FALDs could be markedly different.[14]

Figure 1.2 shows a modification of the model in figure 1.1 to incorporate PALDs in the assessment process. As illustrated in figure 1.2, FALDs are influenced by three different factors: the framework, PALDs, and the assessment (both the exercises and the examinee's responses to these exercises).

Achievement level descriptions are influenced by three different factors: the framework, preliminary descriptions, and the assessment …

The possibility that the two sets of achievement level descriptions may differ limits the usefulness of PALDs as a guide for exercise development. In fact, Lazer, Campbell, and Donahue

---

[11] The 1996 Science Assessment was originally scheduled for 1994 but did not occur until 1996. The PALDs for this assessment were added after the frameworks and specifications were completed.

[12] The frameworks for both the 1996 and 1998 assessments contain similar statements.

[13] The use of noneducators as part of the achievement levels panels has been questioned by some measurement experts. For example, Mehrens (1995, pp. 246–247) writes:

> I do not see how the general public could make decisions about what fourth or eighth graders should know before being promoted—or even classified as advanced, proficient, or basic. I see *some* logic in the general public being represented on a panel setting a standard on what a high school graduate should know or be able to do [either to graduate or simply to be classified]. However, from a purely methodological point of view (ignoring politics), I would always prefer the panel to be experts on the domain being assessed and, if children are involved, on the developmental level of the children being assessed.

[14] For the U.S. History and Geography Assessments, the two sets of descriptions seem fairly similar.

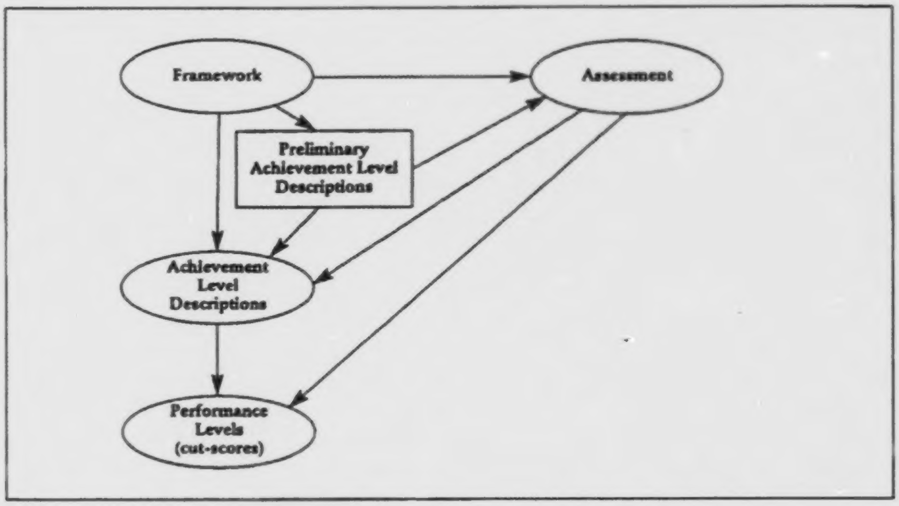

Framework
Assessment
Preliminary Achievement Level Descriptions
Achievement Level Descriptions
Performance Levels (cut-scores)

Figure 1.2. Modification of figure 1.1 to include preliminary achievement level descriptions

(1996), in an extensive discussion of the development of NAEP objectives, items, and background questions for the 1994 Reading, U.S. History, and Geography Assessments, do not specifically mention the use of PALDs as part of the exercise-development process. They identify a 15-step procedure used to develop the items, and none of these steps indicate that the exercises were evaluated for their congruence with PALDs. However, in each assessment, the development, review, and selection of exercises were guided by an Instrument Development Committee (National Center for Education Statistics, 1996d). Because several members of the Framework Planning Committee were also members of this Instrument Development Committee, PALDs for U.S. History and Geography could have been systematically considered throughout the item-development process.[15] Of course, PALDs might serve their most important role as initial statements of what content experts consider to be reasonable achievement levels. Nonetheless, additional information on the role of PALDs in the development of the exercise pool would seem an important part of an overall evaluation of the usefulness of PALDs.

Selected Framework Characteristics and Achievement Levels

This section discusses three characteristics of the frameworks (exercise format, breadth of the content dimension, and

---

[15] The National Academy of Education (1993, p. 123) recommends the establishment of "standing subject-matter panels" and described the functions of such panels as follows:

> The panels should provide continuity to the assessment by being involved in all aspects of the process, including formulating the framework and objectives; reviewing items, item-scoring rubrics, and reporting formats; and helping to achieve agreement on narrative descriptions of performance standards and representative illustrative tasks.

The members of the Instrument Development Committee who were also members of the Framework Planning Committee seem to be performing some of these functions. It is not clear, however, whether the members also provide input to the subsequent components of the assessment process.

Table 1.1. Achievement level cut-scores for U.S. History and World Geography, separately for dichotomous and partial-credit items

Achievement Levels/Item Types	Grade 4	Grade 8	Grade 12
World Geography			
Basic			
Dichotomous Items	182	230	243
Partial-Credit Items	188	247	272
Proficient			
Dichotomous Items	236	275	295
Partial-Credit Items	244	291	313
Advanced			
Dichotomous Items	271	306	329
Partial-Credit Items	286	330	350
U.S. History			
Basic			
Dichotomous Items	171	226	264
Partial-Credit Items	200	261	303
Proficient			
Dichotomous Items	239	282	315
Partial-Credit Items	246	302	334
Advanced			
Dichotomous Items	272	321	346
Partial-Credit Items	283	334	365

Source: National Academy of Education, 1996, p. 104

clarity and complexity of the cognitive dimension). Their impact on the achievement level descriptions or the performance levels (cut-scores) is also considered.

Exercise Format

As noted above, the frameworks provide specific guidelines both for the exercise formats that are to be used (e.g., multiple-choice items and extended responses) and for the distribution of testing time allocated to these formats. For many assessments, detailed item writing guidelines are supplied. For example, the U.S. History specifications list six requirements that alternatives for the multiple-choice items should meet (National Assessment Governing Board, 1992, pp. 16–17).

One of the most pervasive findings from analyses of recent NAEP data is the consistently lower achievement level cut-scores set for items scored dichotomously relative to the cut-scores based on items using a multiple-score scale and permitting partial credit. Examples of the differences in the cut-scores associated with the two item types are shown in table 1.1 for the 1994 U.S. History and Geography Assessments. Results similar to those shown in table 1.1 have also been observed in other assessments.[16] Various hypotheses have been proposed to explain the differences. For example, Shepard (1995) finds that flaws in the standard-setting methodology are the primary cause. Kane (1995) raises other possibilities related to scaling and dimensionality problems.

---

[16] Such cut-score differences would be particularly critical if the item specifications for a given assessment area were to change and if comparisons of the results of the earlier assessment with the results of the new assessment were to be made.

Results such as those shown in table 1.1 are disconcerting because, as the National Academy of Education (1996, p. 93) observes, item features (e.g., format, difficulty, and number of points used in the scoring rubric) should be "irrelevant for cut-score determinations." Because these features should be irrelevant to both the setting of the cut-scores and the writing of the achievement level descriptions, the implications of these findings for developing the frameworks and for the role the frameworks play in the achievement level-setting process would seem to be minimal. Even if the causes of the differences in cut-scores are known, should they have an impact on the frameworks? For example, should the testing time allocated to multiple-choice items change because of these cut-score differences? The primary purpose of the frameworks is to enable the development of an assessment that provides the best representation of the achievement of important educational outcomes in a domain. Accomplishing this purpose should not be influenced by data obtained later in the assessment process.[17]

the frameworks have received considerable praise for their definition of the content domain.

Breadth of the Content Dimension

As indicated in the first part of this paper, the frameworks have received considerable praise for their definition of the content domain. However, their role as the "pivotal link" between the assessment and the achievement level descriptions has not been extensively evaluated. In this role, the frameworks guide the writing of PALDs and ultimately, along with the assessment results, the writing of FALDs and the setting of the performance levels.

As noted previously, achievement level descriptions indicate what students at these levels should be able to accomplish. As such, they represent what are labeled "criterion-referenced (CR) interpretations" of performance. Millman has observed that well-defined domains alone are "insufficient" to guarantee reasonable CR interpretations. In an article reviewing the history of CR testing, he writes:

> Clear and well-explicated domains are insufficient to assure interpretability. If the domain defines a broad construct—such as, knowledge of the American Civil War—no matter how well spelled out it is, with a limited number of test items, we still won't know what tasks within that domain the student can and cannot do. We can construct reading proficiency and mathematical reasoning scales. We can place students on such scales, a highly important measurement function. However, we would probably still not know what tasks the student can and cannot do. Low task intercorrelations—that is, task specificity—work against such CR interpretations. Reporting by a narrower domain, such as the Battle of Gettysburg, helps only if enough items are sampled from that domain. (Millman, 1994, pp. 19–20)

Millman also considers the nature of NAEP domains and the possibility of CR interpretations with such domains:

---

[17] This statement does not mean that such differences should be ignored at all steps in this process. Such results may have implications for the frameworks of future assessments in an area.

> NAEP tests are just not designed to provide, nor do they claim to provide, the promised CR interpretation. Their constructs are too broad. ... Their role as "The Nation's Report Card" requires, for all practical purposes, that progress be reported in broadly defined domains.
>
> Two recent changes in NAEP's operation have been the return of attention to performance assessment and the use of the categories—Basic, Proficient, and Advanced—to report levels of achievement. Will each of these two shifts add to the CR interpretability of NAEP results? The answers are no and no. (Millman, 1994, pp. 20 and 39)

To illustrate Millman's concern about "broadly defined domains," the specifications of the 1994 Geography and Reading Assessments are considered below.

Detailed content specifications in the 1994 Geography Framework (National Assessment Governing Board, 1994a) provide lists of statements identified as the "Content Outline."[18] The number of such statements across the three main content categories (space/place, environment/society, and spatial dynamics/connections) and the number of exercises in the final assessment (National Center for Education Statistics, 1996d) are shown in Table 1.2 below.

Given this information, the coverage of the domain by the assessment would still be somewhat limited at each grade level, even if each statement could be measured directly by a single item. However, the number of potential exercises associated with a given statement varies considerably. Some statements seem to require only a single item to measure adequately the implied learning target [e.g., knowing the difference between fertile and infertile soils] (National Center for Education Statistics, 1996a, p. 35).[19] For other statements, a large number of exercises could be written, and the number of exercises required to measure the learning target adequately might be difficult to determine. Consider, for example, the following two statements:

1. Use great circle routes to measure distances on a globe. (National Center for Education Statistics, 1996a, p. 35)
2. Understand how patterns and processes in human geography are interrelated in the world, such as how the growth in the number of immigrants often leads to an increasing number of minority groups in a country. (National Center for Education Statistics, 1996a, p. 38)

Table 1.2. 1994 Geography Assessment

Grade	Number of Statements in the Content Outline	Number of Exercises in the Assessment
4	191	90
8	195	125
12	164	123
Total	550	338*

*Some exercises were administered at two grade levels. A total of 273 unique exercises were administered.

---

[18] In the context of classroom instruction, these statements would probably be considered "learning targets" or "learning objectives" (Nitko, 1996).

[19] All geography statements are taken from the grade 4 content outline.

Clearly, a very large number of exercises could be written to measure the content of the first statement. However, the responses to these exercises would probably be so highly correlated that it might be possible to say that a student could or could not perform this learning target on the basis of just one or two exercises.[20,21] Highly correlated means that if a student answers one exercise correctly, then he/she would "very likely" answer the other exercises correctly.

Again, for the second statement, a very large number of exercises could be written. One of these could be related to the example given as the last part of the statement. The total number of other possible relationships in human geography that could be included as part of this content is difficult to imagine. Furthermore, a student might "understand" some relationships but not others.[22] In other words, if a student answers one exercise correctly, it is not necessarily "highly likely" that he/she would answer the other exercises correctly. (To illustrate this type of situation, specific data are reported below as part of the material related to the breadth of the NAEP Reading Assessment domain.)

initial
understanding
and developing
an interpretation
represent two . . .
reading behaviors . . .

The NAEP Reading Assessment provides another illustration of Millman's concern about the breadth of NAEP domains.

The assessment/exercise specifications for the 1994 Reading Assessment (National Assessment Governing Board, 1990) differ markedly from the 1994 Geography Assessment Specifications, most importantly because no detailed content outline for the Reading Assessment is included.[23] The content dimension in the Reading Framework is divided into three categories: reading for literary experience, reading to be informed, and reading to perform a task. Although no content outline is given, the number of potential stimuli for each domain is enormous. Consider, for example, the reading for literary experience domain. How many "fantasies, fables, fairy tales, myths, mysteries, realistic fiction, adventure stories" could have been identified as possible reading passages for the 1994 Reading Assessment (National Assessment Governing Board, 1990, p. 3)? Obviously, the supply of such material is virtually unlimited, and, therefore, the number of possible exercises is also unlimited.

However, as noted in the discussion of the geography domain, a large number of exercises do not necessarily imply that CR interpretations would be difficult to make. If the exercises based on different passages exhibit little task specificity (i.e., if the exercises are highly correlated), then CR interpretations might still be reasonable.

For the Reading Assessment, "initial understanding" and "developing an interpretation" represent two of the four reading behaviors that are to be assessed

---

[20] If multiple-choice items were used and, therefore, the possibility of guessing the correct answer was a factor, additional exercises may be desirable. Actually, even if multiple-choice items were not used, measurement errors would still have to be considered.

[21] If generalizations across different sizes of globes with different scales is a concern, then other exercises might be needed. However, if the concern is strictly with the procedure for using the great circle routes and not with the arithmetic, then a judgment about the mastery of the procedure might be based on one or two exercises.

[22] In this discussion of the second statement, the problems faced by item writers when they attempt to operationalize "understand how" are not considered. These problems are considered in the next section.

[23] Detailed content outlines exist for the U.S. History, Science, and Civics Assessments. However, the Writing Assessment, like the Reading Assessment, does not have such an outline. Of course, given the "process nature" of these two domains, the lack of a detailed content outline would probably not be considered a problem.

regardless of the content category (National Assessment Governing Board, 1990). These two behaviors are defined as follows:

> *Initial understanding* requires a broad, preliminary construction of an understanding of the text. Questions testing this aspect should ask the reader to provide an initial impression or unreflected understanding of what was read. The first question following any passage should be one testing initial understanding.
>
> *Developing an interpretation* requires the reader to go beyond the initial impression to develop a more complete understanding of what was read. Questions testing this aspect should require a more specific understanding of the text and involve linking information across parts of the text as well as focusing on specific information. (National Assessment Governing Board, 1990, p. 10)

Would questions measuring each behavior be highly correlated over passages? To provide an answer to this question, data from two sources are presented. One source deals with exercises related to "developing an interpretation behavior" and the other with exercises focused on "initial understanding behavior."

In an earlier paper, I observed the following concerning exercises that would be classified in the "developing an interpretation" category:

> As an illustration of this problem [low correlation between exercises], consider the following questions taken from the *Iowa Tests of Educational Development:*
>
> 1. From his manner and formal training, what opinion might people have formed of John Marshall? (28%)
> 2. What do the last two sentences suggest about Patasonians' acceptance of U.S. aid? (44%)
> 3. Suppose an uninsured and unemployed motorist damaged someone's car. Which speaker offers a plan that would allow the injured party to collect benefits? (64%)
>
> Each of these items is associated with a particular reading passage, and all three items require the student to reach a conclusion on the basis of the information in the passage. The percentage in parentheses after each question show the percentage of 10th grade students from a statewide sample in Iowa who answered the item correctly. The varying percentages associated with the three questions above provide evidence that getting one item correct does not guarantee that a second item measuring the same objective (where the objective is defined in fairly broad terms) will be answered correctly. Furthermore, though this is not discernible from the data given above, all 28% who got the first item correct did not get the second or third items correct (Forsyth, 1976, pp. 12–13).

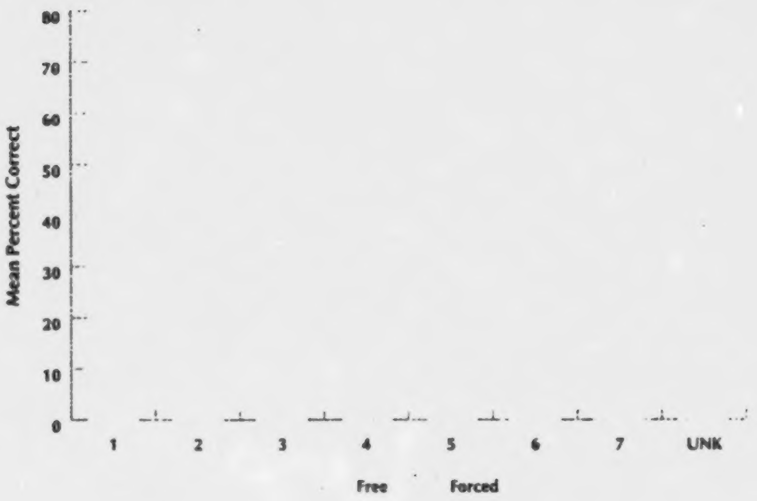
Would questions measuring each behavior be highly correlated over passages?

The second source of information related to task specificity is a study by Allen and Isham (1996).[24] Allen and Isham present data for grade 8 examinees that were collected as part of a special study (based on the *NAEP Reader*) in the 1994 NAEP Reading Assessment.

---

[24] This study represents a unique investigation of the task specificity issue.

Figure 1.3. Mean scores for item 1



Source: Allen & Isham, 1996, p. 13

The purpose of their study was:

> to verify the assumptions that are made by those who are proponents of the use of choice in the *NAEP Reader*. These assumptions are that the generic questions have the same meaning no matter what story was read, and that students do best when they are able to read a story that they are interested in, one that they select. (National Center for Education Statistics, 1996d, p. 3)

Only information related to the assumption that the "generic questions have the same meaning" is of interest at this time. The basic data related to this issue were gathered by having each of seven randomly equivalent groups of eighth-grade examinees respond to a different reading passage. After reading these different passages, all examinees answered the same set of generic questions. Although Allen and Isham do not provide the specific set of generic questions, they provide the following definition: "An example of a generic question is one that asks about the appropriateness of the title of the story" (Allen & Isham, 1996, p. 3).[25] All the generic questions were the constructed-response type.

To illustrate the outcomes of this study, the data related to the first item (a dichotomously scored item) in the set of 11 generic items are discussed.[26] Figure 1.3 shows the mean scores (percentage correct) of the seven representative groups of eighth-grade examinees for the first item.[27] The figure shows that these percentage correct values range from approximately 40 percent for story 4 to approximately 80 percent for story 3. Allen and Isham note that the

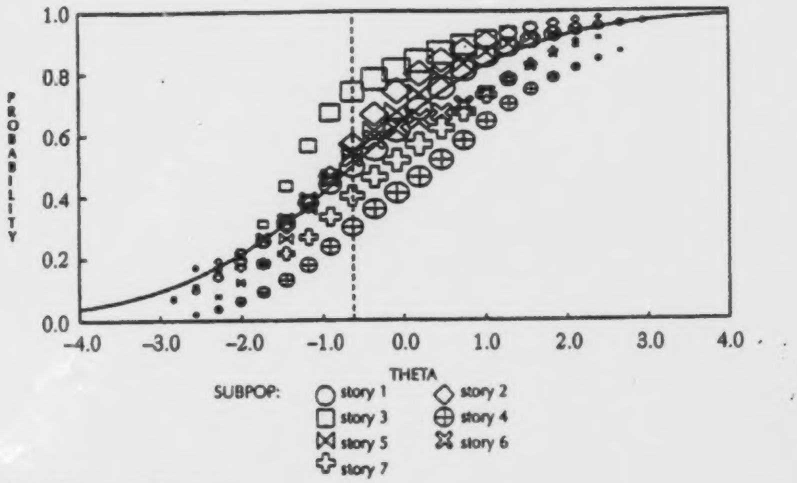
---

[25] How the question was stated is not known. One possibility: Is the title of this story appropriate? Briefly explain your answer. A second possibility: What is an appropriate title for this story? Briefly explain your answer.

[26] If the specifications for the *NAEP Reader* passages were the same as those for the Reading Assessment, this first item tested "initial understanding." Perhaps this item asked about the appropriateness of the title.

[27] The figure also shows the mean scores for seven groups of students who were allowed to select their reading passages. These means do not pertain to this discussion.

Figure 1.4. Item response functions for item 1



Source: Allen & Isham, 1996, p. 13

differences among the groups are "quite striking" (Allen & Isham, 1996, p. 8).

To provide additional information about the comparability of generic questions across reading passages, each generic question was scaled using item response theory "as if [the question] had the same meaning no matter which story was used" (Allen & Isham, 1996, p. 8). Figure 1.4 shows the item response functions (IRF) for item 1. As Allen and Isham observe, these IRFs differ across the range of the proficiency scale.

One way to examine the possible implications of the results shown in figure 1.4 is to speculate about how these "identical items" might be used in the reporting of NAEP results. The 1994 NAEP Report Cards (National Center for Education Statistics, 1996a, 1996b, 1996c) report NAEP results using three different procedures: scale anchoring, item mapping, and achievement levels.[28] Figure 1.5 is adapted from the *NAEP 1994 Reading Report Card* (National Center for Education Statistics, 1996b, p. 93) and shows the mapping of selected items on to the reading for literary experience subscale. The question is: Where would generic item 1 map on to this scale? Given the results in figure 1.4, the answer depends on which story (stories) had been included in the assessment. To illustrate this point, assume that the mean and standard deviation of the literary experience subscale are 259 and 37, respectively.[29] Given these values and using the same item-mapping procedure employed with the NAEP data,

---

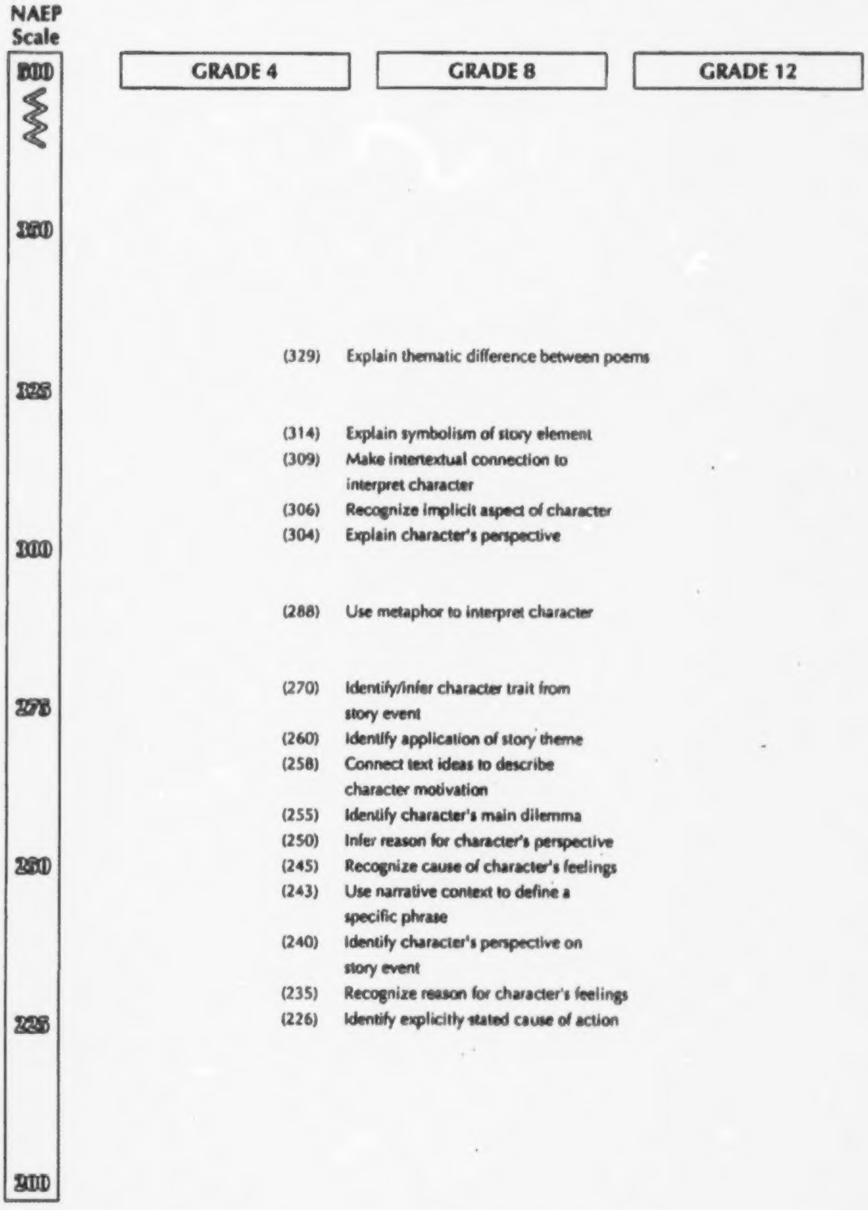
[28] The scale-anchoring and item-mapping procedures are described in Appendix B of the *NAEP 1994 Reading Report Card for the Nation and the States* (National Center for Education Statistics, 1996b).

[29] The mean of this subscale is reported as 259 in the *NAEP 1994 Reading Report Card for the Nation and the States* (National Center for Education Statistics, 1996b, p. 87). However, a standard deviation value for this subscale could not be located. Therefore, the standard deviation for the composite reading scale was selected (ibid, 1996b, p. 307).

Figure 1.5. Mapping of reading items

Figure 6.4. Map of Selected Items on the Reading for Literary Experience subscale for Grades 4, 8, and 12

Each reading question was mapped onto the NAEP literary subscale based on students' performance. The point on the subscale at which a question is positioned on the map represents the subscale score attained by students who had a 65 percent probability of successfully answering the question. Thus, it can be said for each question and its corresponding subscale score—students with proficiency scores above that point on the subscale have a greater than 65 percent chance of successfully answering the question, while those below that point on the subscale have a less than 65 percent chance. (The probability was set at 74 percent for multiple-choice questions.) In interpreting the item map information it should be kept in mind that students at different grades demonstrated these reading abilities with grade-appropriate reading materials.



Source: Adapted from National Center for Education Statistics, 1996b, p. 94

the following estimated scale values would be associated with stories 1, 3, and 4:[30]

Story		Scale Value
3	—	222
1	—	259
4	—	296

Assume that the descriptor for this first item is: Recognize main topic.[31] In figure 1.5, the "Recognize main topic" descriptor would be near the top of the map, in the middle of the map, and at the bottom of the map. Given the purpose of the item-mapping procedure, such an outcome would be somewhat confusing.[32,33]

Not all of the generic questions exhibited the pattern of results shown in figures 1.3 and 1.4. However, based on their evaluations of the 11 generic questions, Allen and Isham conclude:

> In summary, some of the generic questions seem to have similar characteristics no matter which story they refer to, but others have very different characteristics depending upon the story. These similar and differing characteristics are reflected in mean item scores and in empirical item response functions. *Given that the generic questions do not seem to have the same characteristics across the seven stories, treating the questions as being the same no matter which story was read is inappropriate* [italics added]. (Allen & Isham, 1996, p. 8)[34]

This extensive discussion of the Geography and Reading Frameworks was intended to reinforce Millman's observation that NAEP domains are "broadly defined" and to illustrate the problems such domains create when CR interpretations are attempted. It is important to note that these problems have an impact on any reporting system that attempts to provide CR interpretations for NAEP results. (This was Millman's point also.) Thus, scale-anchoring, item-mapping, and achievement level procedures must all deal with these problems in some way.

". . . treating the questions as being the same no matter which story was read is inappropriate. . . ."

Most of the interpretation problems that reporting procedures encounter are created when a single item is used to represent a construct for which it is

---

[30] These values are relatively crude estimates. However, using a probability value of 0.65 (vertical axis), the q value associated with this probability is approximately -1 (one standard deviation below the mean) for story 3 and approximately +1 (one standard deviation above the mean) for story 4.

[31] Items with such a descriptor would seem to be in the "initial understanding" category.

[32] This type of situation already exists to some extent in the item map shown in figure 1.5. Consider the following two descriptors:

(245)	Recognize cause of character's feelings
(235)	Recognize reason for character's feelings

Likewise, understanding the difference between the behaviors represented by the following two descriptors might be difficult for most people:

(306)	Recognize implicit aspect of character
(235)	Recognize reason for character's feelings

[33] In this example, the three items had scale values from low to high. It could be argued that in this situation only the item with the lowest value should have been included in the map. However, including only this item would be misleading since examinees with proficiency levels at the lower part of the scale could recognize the main topic in only one out of seven reading passages. Other interpretation issues would be raised in different situations. For example, assume that only story 4 was part of the assessment. The only place "Recognize main topic" would appear would be at the top part of the map. In such a situation, it would be assumed that examinees with proficiency levels in the middle of the scale could not recognize main topics. Such an outcome would seem inconsistent with other descriptors in the middle of the scale (e.g., Identify application of story theme).

[34] The importance of this conclusion cannot be overstated, even though it is derived from the results of a single study.

an inadequate representation.[35] Of the three reporting procedures previously noted, the interpretation burden placed on a single item seems greatest for the item-mapping procedure, as illustrated by the above discussion. However, similar concerns occur with the scale-anchoring procedure because, once again, the descriptions of what examinees can accomplish are based on the results from specific items. To illustrate this concern, consider the results reported in the *NAEP 1994 Reading Report Card for the Nation and the States* (National Center for Education Statistics, 1996b).

Of course, for NAEP reports, many subdomains are combined before general achievement level statements are made.

Scale anchoring is used to anchor the reading composite proficiency scale at the 25th, 50th, and 90th percentiles. The scale value for the 25th percentile is 236 (National Center for Education Statistics, 1996b, p. 23). One descriptor for this percentile is "recognize main topics" (p. 85). Assume that generic item 1 in the Allen and Isham study was related to this outcome. What if only stories 1 and 4 had been included in the Reading Assessment? Given the criteria used to identify possible anchoring items, it is highly unlikely that these two items would be available to help describe what examinees near the 25th percentile can accomplish. Under these conditions, the anchor descriptions probably would have changed.[36]

The achievement levels procedure would seem to be less susceptible to this particular interpretation problem because the descriptions should not be as dependent on individual items as are the other two procedures.[37] However, it would seem reasonable for these descriptions to recognize specifically the limitations placed on the ability to make unqualified statements about what students can accomplish. As the above examples illustrate, large numbers of exercises could be written, even for what might be considered very narrow subdomains such as "recognize main topics" or "understand how patterns and processes in human geography are interrelated in the world," and the interrelations among the exercises within the subdomain may be low. Of course, for NAEP reports, many subdomains are combined before general achievement level statements are made. Given such conditions, perhaps the achievement level descriptions (both PALDs and FALDs) should recognize that as the achievement level increases, the frequency with which examinees either can perform certain behaviors or know the information in certain domains also increases. One illustration of statements describing such a relationship, taken from fourth-grade PALDs of the U.S. History Framework (National

---

[35] In a slightly different context, Stone (1995) discusses a "single item syndrome." Stone writes: "A single item may not be an adequate representation of a body of knowledge and by placing such emphasis on the sanctity of the single item, disturbing results may evidence themselves" (p. 12). Lissitz and Bourque (1995) make a similar point when they wrote that "describing what [individual items] mean on an ability continuum (the NAEP scale) is a high-inference task" (p. 17).

[36] As explained in appendix B of the *NAEP 1994 Reading Report Card for the Nation and the States* (National Center for Education Statistics, 1996b), the anchoring process is considerably more complicated than this simple example indicates. However, the example illustrates an important issue for scale anchoring.

[37] Reckase (1993) and Lissitz and Bourque (1995) make similar observations about achievement level procedures relative to scale-anchoring procedures.

Assessment Governing Board, 1994e, p. 49), is shown below:

> Basic ... Should be able to identify and describe *a few* of the most familiar people, events, and documents in American history.
>
> Proficient ... Should demonstrate familiarity with *a number* of historical people, places, and events.
>
> Advanced ... Should demonstrate *a considerable* familiarity with historical people, places, and events.[38]

Of course, standard-setting panel members would face the difficult task of using both these frequency indicators and their knowledge and understanding of the content domain to arrive at the specific performance levels on the NAEP scale.

Clarity and Complexity of the Cognitive Dimension

The cognitive dimension forms an important part of most frameworks.[39] The purpose of this dimension is to ensure that the entire spectrum of student outcomes is represented in the exercise pool. For example, the geography framework uses three cognitive categories: knowing, understanding, and applying.[40] Each item in the exercise pool is classified into one of these categories. However, Lazer, Campbell, and Donahue observe that such classifications might not be very accurate.[41] Concerning the geography classifications, they state:

> It should be noted that the classification of items into different cognitive categories—conducted by both Educational Testing Service staff and members of the assessment development committee—is likely an imprecise process. (Lazer, Campbell, and Donahue, 1996, p. 62)

Similar statements accompany discussion of the development of the U.S. history (Lazer, Campbell, & Donahue, 1996, p. 53) and reading (p. 43) exercises. Given that *The NAEP 1994 Technical Report* (National Center for Education Statistics, 1996d) provides no data related to the magnitude of the imprecision in these classifications, formal procedures for investigating this concern were evidently not undertaken.[42,43]

However, the National Academy of Education investigated the consistency of the cognitive classifications between NAEP item developers and outside experts for the 1992 Mathematics Assessment (grade 4) and the 1992 and 1994 Reading Assessments (grades 4, 8, and 12). Tables 1.3 and 1.4 present some of the results from these investigations.[44] For the 157 items in the 1992 fourth-grade Mathematics Assessment, the two groups of raters agreed on the cognitive classifications for 109 (69.4%) items

---

[38] These statements are from the PALDs. The Proficient and Advanced statements were changed in the FALDs. For Proficient, "many" was used in place of "a number of" and for Advanced this statement was omitted.

[39] The 1998 Writing Assessment does not have a formal cognitive dimension.

[40] Applying is a very broad category. "[It] involves the higher-order thinking processes of classifying, hypothesizing, using inductive and deductive reasoning, and forming problem-solving models" (National Assessment Governing Board, 1994a, p. 8).

[41] The accuracy of the classifications in the content categories is usually not questioned.

[42] Sireci (undated, p. 57) notes that "an entire literature exists documenting procedures available for determining how well the items comprising a test matches [its] content and cognitive specifications. . . . However, none of these procedures [has] been applied to NAEP tests!"

[43] Possible explanations for this imprecision were not provided.

[44] The results in table 1.3 are for the 1994 Reading Assessment. Similar results were observed for the 1992 Reading Assessment (National Academy of Education, 1993b, p. 67).

Table 1.3. Process classification of 1992 fourth-grade Trial Assessment Mathematics items: Cross-classified by NAEP and by expert raters

NAEP	Expert Raters			Toal Items as Judged by NAEP
	Conceptual Understanding	Procedural Knowledge	Problem Solving	
Conceptual Understanding	38 (58%)	15 (23%)	13 (20%)	66
Procedural Knowledge	1	25 (78%)	6	32
Problem Solving	6	7	46 (78%)	59
Total Items as Judged by Expert Raters	45	47	65	157

Source: National Academy of Education, 1993b, p. 61

Table 1.4. Classification of 1994 Reading Assessment items: Expert advisor classifications compared with official NAEP classifications (grades 4, 8, and 12 combined)

Expert Advisor Classification	Official NAEP classifications				TOTAL N % of Column Total
	Initial Understanding N % of Column Total	Developing Interpretation N % of Column Total	Personal Response N % of Column Total	Critical Stance N % of Column Total	
Initial Understanding	20 71%	24 16%	1 2%	4 4%	49 14%
Developing Interpretation	7 25%	120 79%	9 16%	48 46%	184 54%
Personal Response	NA	2 1%	39 71%	4 4%	45 13%
Critical Stance	1 4%	5 3%	6 11%	48 46%	60 18%
TOTAL	28	151	55	104	338

Source: National Academy of Education, 1996, p. 24

(table 1.3). Of the 66 items classified as conceptual understanding by NAEP, 57.6% were so classified by external raters. Fifteen (22.7%) of these 66 items were classified as procedural knowledge, and 13 (19.7%) were classified as problem solving. For the 338 reading items (all grades) in the 1994 Reading Assessment, the two groups of raters agreed on the cognitive classifications for 227 (67.2%) items (table 1.4). Less than half (46.2%) of the items that NAEP classified as critical stance items received a similar classification from the external raters. The external raters also classified 46.2% of the items that NAEP classified in the critical stance category as developing interpretation items.

Data similar to those in tables 1.3 and 1.4 are not available for recent assessments (e.g., 1994 U.S. History and Geography Assessments). However, given the mathematics and reading results and the observations by Lazer, Campbell, and Donahue (1996), considerable rater disagreement with respect to the cognitive classifications seems likely. In content areas such as U.S. history and geography, one possible reason for the imprecision in the classification process is related to the interaction of examinees' past experiences and the content of the item. Consider, for example, a content statement from the geography content outline noted previously: "Know the difference between fertile and infertile soils." Assume that the item writer decides to measure this learning target using the question: What is the difference between fertile and infertile soils?[45] For those fourth-grade students who have encountered discussions of the difference between the two types of soil and who remember that discussion, this item would belong in the "knowing" category. However, for students who have not encountered such a discussion, this item might require the use of some of the higher-order skills in the "applying" category so they could answer the item correctly. The same issue surfaces when trying to distinguish between the "knowing" and "understanding" categories. Consider, for example, the question given in the 1994 Geography Framework (National Assessment Governing Board, 1994b, p. 12) to illustrate the "understanding" category: "Why are tropical rain forests located near the equator?" If students have been taught this generalization, is it a "knowing" question or an "understanding" question?[46]

Underlying this issue is the premise that some new elements must be included in the item before it can be labeled as measuring one of the "higher" cognitive categories. More than 50 years ago, the National Society for the Study of Education (1946), in a book devoted to the measurement of understanding, addressed this issue in the following way:

> *The Need for Novelty.* Much that passes for understanding in the school is primarily memorization. On the other hand, understanding is attested when the pupil dips into his knowledge and fits it into new patterns of thought or action which could not have been directly learned. For example, being able to recite a number of reasons why something happened is no real assurance that the person comprehends *why* it happened. He may not understand (sense the significance of) the reasons that he has learned. It is no more difficult for the pupil to learn reasons as facts than to learn names and dates as facts. Any genuine test of understanding will, therefore, require that the pupil show his ability to utilize knowledge (perhaps of relationships) to explain or interpret events in new situations or contexts. Accordingly, understanding should not emphasize

The same issue surfaces when trying to distinguish between the "knowing" and "understanding" categories.

---

[45] Although this question may lack creativity, it matches the content statement.

[46] In some content areas, this problem is exacerbated because most schools do not have specific classes for these subjects—consider, for example, fourth-grade geography and U.S. history.

> reasons or supposed insights which have been taught and learned as facts, but should call for the use of abilities, both detailed and general, to cope with situations containing at least some novel elements. (National Society for the Study of Education, 1996, p. 40)

Clearly, this "need for novelty" requires persons who classify items in cognitive categories to make assumptions about both the background knowledge needed to answer an item and the experiences of the typical examinee regarding that knowledge. Thus, different classifications of the items will occur, because not all people will make the same assumptions.

Including a significant number of specific examples in the higher-level cognitive categories should help item development

Given these concerns about novelty, it would be helpful both to the item developers and to the standard-setting panels if increased numbers of concrete examples were used in the frameworks to illustrate what the content experts view as representing the "higher" cognitive categories. Most of the examples provided as part of the statements in the content outlines are not adequate for defining the cognitive categories clearly. Consider, for example, the following statement from the fourth-grade geography outline:

> Analyze the processes that shape cultural patterns, cause trends in population growth, and/or influence travel destinations; for example, the discovery of coal and oil in western Pennsylvania led to the development of the industrial area around Pittsburgh. (National Assessment Governing Board, 1994a, p. 56)

Presumably, the content experts want items related to this statement to be classified in the "applying" category.[47] However, the example put forth provides little if any guidance concerning what "analyze" means and how these analysis behaviors should be measured. Including a significant number of specific examples in the higher-level cognitive categories in the frameworks should help item development as well as increase rater agreement with respect to the item classifications in these categories. In addition, standard-setting panels would benefit from the increased clarity such examples would provide for the cognitive dimensions.

The cognitive categories generally represent increasing levels of complexity in thinking processes.[48] Furthermore, these processes and the general content categories are usually the same for all grade levels. Thus, if achievement level descriptions reflect what students should be able to do with grade-appropriate material,[49] these descriptions should probably exhibit some similarity across grade levels. For example, the achievement level descriptions for Basic should be similar regardless of grade level. In fact, current achievement level descriptions for the Reading Assessment exhibit some similarities, as the statements below illustrate:

---

[47] See footnote 41.

[48] For the 1992 and 1994 Reading Assessments, these categories "do not form a sequential hierarchy" (National Assessment Governing Board, 1994c, p. 13).

[49] Most Report Cards remind readers that NAEP results are based on "grade-appropriate" materials (see figure 1.5). Of course, the definition of grade-appropriate material presents its own set of problems. These problems will probably be particularly complex when specific grade-level courses typically do not exist (e.g., grade 4 geography).

Grade 4 Basic. Fourth-grade students performing at the Basic level should:

- Demonstrate an understanding of the overall meaning of what they read.
- Be able to make relatively obvious connections between the text and their own experiences.
- Extend the ideas in the text by making simple inferences.

Grade 8 Basic. Eighth-grade students performing at the Basic level should:

- Demonstrate a literal understanding of what they read and be able to make some interpretations.
- Be able to identify specific aspects of the text that reflect the overall meaning and extend the ideas in the text by making simple inferences.
- Recognize and relate interpretations and connections among ideas in the text to personal experience.
- Draw conclusions based on text.

Grade 12 Basic. Twelfth-grade students performing at the Basic level should:

- Be able to demonstrate an overall understanding and make some interpretations of the text.
- Be able to identify and relate aspects of the text to its overall meaning, extend the ideas in the text by making simple inferences, and recognize interpretations.
- Make connections among and relate ideas in the text to their personal experiences.
- Draw conclusions.
- Be able to identify elements of an author's style. (National Center for Education Statistics, 1996b, p. 42)

Two observations about these reading descriptions seem relevant. First, the descriptions are linked more to the four cognitive categories (initial understanding, developing an interpretation, personal reflection and response, and critical stance) than to the three content categories (literary experience, gain information, and perform a task). Second, an implied frequency/complexity dimension is present in these statements. Consider fourth-grade Basic. In the domain of inferences, these students should make "*simple* inferences." In the domain of personal reflection and response, these students should make "*relatively obvious* connections." In these statements, the implied frequency dimension is linked to the complexity of the questions the student was asked to answer (simple inferences, relatively obvious connections).[50]

. . . the (reading) descriptions are linked more to the four cognitive categories . . . than to the three content categories . . .

The frequency/complexity link is not part of the Proficient descriptions for grades 4, 8, and 12. Instead, these descriptions (at all grade levels) include the phrase "extend the ideas in text by making inferences." If "simple" were used as a qualifier in the description of Basic (i.e., indicating a subset of all items), then the absence of any qualifier in the Proficient statement would seem to indicate that Proficient students should be expected to make "all" inferences.[51]

---

[50] In the previous example from U.S. history, the frequency dimension was linked to the number of facts (people, places, and events) the examinee should know.

The above example raises the question: How applicable to other content areas is the model of achievement level descriptions represented by the reading assessment descriptions? For example, would this model be useful in developing PALDs in civics or writing?[52] The fundamental premise of such a model seems to be that, whereas the content categories (or themes) remain constant across grade levels, the learning targets (i.e., grade-appropriate content) become more comprehensive and more complex.

Concluding Statement

The NAEP frameworks have received considerable praise both for the process used to develop them and for their comprehensive coverage. This paper has examined the potential influence of a few select framework characteristics on achievement level descriptions or performance levels (cut-scores). The major conclusions of this examination are:

1. Given the breadth of the NAEP domains, the achievement level descriptions should include an explicit "frequency dimension."
2. Given the lack of clarity of the cognitive dimension of the frameworks, both item developers and standard-setting panels would benefit from increasing the number of concrete examples in the frameworks to illustrate how content experts think that higher-level cognitive categories should be measured.
3. Given the nature of the content and cognitive dimensions and the fact that the interpretation of NAEP results assumes grade-appropriate materials, the possible use of achievement level descriptions that are relatively similar across grade levels should be examined.

References

Allen, N. L., & Isham, S. P. (1996). *The 1994 NAEP reader and information about the impact of choice on item response functions.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Forsyth, R. A. (1976). *Describing what Johnny can do.* Iowa City, IA: Iowa Testing Programs.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice, 10* (3), 3–9, 16.

Kane, M. (1993). *Comments on the NAE evaluation of National Assessment Governing Board achievement levels.* Washington, DC: National Assessment Governing Board.

Kane, M. (1995). Examinee-centered vs. task-centered standard setting. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments,* Volume II (pp. 119–141). Washington, DC: National Assessment Governing Board, National Center for Education Statistics.

---

[51] Given, of course, grade-appropriate materials. (See footnote 49.)

[52] Actually, PALDs for writing follow this model (National Assessment Governing Board, 1998b pp. 52–56). However, PALDs for civics are lists of specific content (from five categories) that students should be able to identify, describe, explain, or defend (National Assessment Governing Board, 1998a, pp. 84–89).

Lazer, S., Campbell, J. R., & Donahue, P. (1996). Developing the NAEP objectives, items, and background questions for the 1994 assessments of reading, U.S. history, and geography. *The NAEP 1994 Technical Report* (pp. 29–67). Washington, DC: National Center for Education Statistics.

Linn, R. L., & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29,* 177–194.

Lissitz, R. W., & Bourque, M. L. (1995). Reporting NAEP results using standards. *Educational Measurement: Issues and Practices, 14* (2), 14–23, 31.

Mehrens, W. A. (1995). Methodology issues in standard setting for educational exams. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments,* Volume II (pp. 221–263). Washington, DC: National Assessment Governing Board, National Center for Education Statistics.

Millman, J. (1994). Criterion-referenced testing 30 years later: Promise broken, promise kept. *Educational Measurement: Issues and Practice, 13* (4), 19–20, 39.

Mullis, I. V. S. (1995). *NAEP: A look at future needs and means.* Washington, DC: National Assessment Governing Board.

National Academy of Education. (1993a). *Setting performance standards for student achievement.* Stanford, CA: Author.

National Academy of Education. (1993b). *The trial state assessment: Prospect and realities.* Stanford, CA: Author.

National Academy of Education. (1996). *Quality and utility: The 1994 trial state assessment in reading.* Stanford, CA: Author.

National Assessment Governing Board. (1990). *Assessment and exercise specifications for the National Assessment of Educational Progress in Reading: 1992–1998.* Washington, DC: Author.

National Assessment Governing Board. (1992). *Assessment and exercise specifications: 1994 National Assessment of Educational Progress in U.S. History.* Washington, DC: Author.

National Assessment Governing Board. (1994a). *Geography assessment and exercise specifications for the 1994 National Assessment of Educational Progress.* Washington, DC: Author.

National Assessment Governing Board. (1994b). *Geography framework for the 1994 National Assessment of Educational Progress.* Washington, DC: Author.

National Assessment Governing Board. (1994c). *Reading framework for the 1992 and 1994 National Assessment of Educational Progress.* Washington, DC: Author.

National Assessment Governing Board. (1994d). *Science assessment and exercise specifications for the 1994 National Assessment of Educational Progress.* Washington, DC: Author.

National Assessment Governing Board. (1994e). *U.S. history framework for the 1994 National Assessment of Educational Progress.* Washington, DC: Author.

National Assessment Governing Board. (1996a). *1998 NAEP civics assessment planning project: Recommended assessment framework and test specifications.* Washington, DC: Author.

National Assessment Governing Board. (1996b). *Science framework for the 1996 National Assessment of Educational Progress.* Washington, DC: Author.

National Assessment Governing Board. (1998a). *Civics assessment framework for the 1998 National Assessment of Educational Progress.* Washington, DC: Author.

National Assessment Governing Board. (1998b). *Writing framework and specifications for the 1998 National Assessment of Educational Progress.* Washington, DC: Author.

National Assessment Governing Board and National Center for Education Statistics. (1995). *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments,* Volume II. Washington, DC: National Assessment Governing Board, National Center for Education Statistics.

National Center for Education Statistics. (1996a). *NAEP 1994 geography report card.* Washington, DC: Author.

National Center for Education Statistics. (1996b). *NAEP 1994 reading report card for the nation and the states.* Washington, DC: Author.

National Center for Education Statistics. (1996c). *NAEP 1994 U.S. history report card.* Washington, DC: Author.

National Center for Education Statistics. (1996d). *The NAEP 1994 technical report.* Washington, DC: Author.

National Center for Education Statistics. (1997). *NAEP 1994 mathematics report card for the nation and the states.* Washington, DC: Author.

National Society for the Study of Education. (1946). *The measurement of understanding: Forty-fifth Yearbook,* Part 1. N. B. Henry, (Ed.). Chicago: University of Chicago Press.

Nitko, A. J. (1996). *Educational assessment of students* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Reckase, M. D. (1993). *The defensibility of domain descriptions developed to support the interpretation of achievement levels and anchor points reported on the NAEP scale.* Iowa City, IA: American College Testing.

Shepard, L. A. (1995). Implications for standard setting of the National Academy of Education evaluation of the National Assessment of Educational Progress achievement levels. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments,* Volume II (pp. 143–160). Washington, DC: National Assessment Governing Board, National Center for Education Statistics.

Sireci, S. G. (undated). *Dimensionality issues related to the National Assessment of Educational Progress.* Washington, DC: National Academy of Sciences.

Stone, G. E. (1995). *Objective standard setting.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

SECTION 2

Assembly of Test Forms for Use in Large-Scale Educational Assessments

Wim J. van der Linden University of Twente, The Netherlands

August 1997

Assembly of Test Forms for Use in Large-Scale Educational Assessments

The problem of assembling test forms for use in large-scale assessments involves treating different statistical features and content constraints for each individual form. This paper outlines two possible methods of test assembly for use in such assessments. In one method, the items are assigned directly from the pool to the test forms. The other method follows the balanced incomplete block design currently in use by the National Assessment of Educational Progress (NAEP).

The practice of educational and psychological testing provides many cases in which the focal point is the assembly of sets of multiple test forms rather than a single form. An obvious example is a testing organization assembling multiple parallel forms of a test for administration at different locations or time slots. Another example is the assembly of test forms for use in a large-scale assessment. The latter example generalizes from the former in that the contents of the individual test forms typically differ. The reason for this is twofold: First, the item pools needed to cover the subject areas assessed are generally too large to administer all items to each student. Second, the populations of students addressed in educational assessments have a structure of several levels of nesting (e.g., classes, schools, districts). Hence, cluster sampling is mostly applied, with higher-level clusters being sampled first, followed by units within these clusters. The negative effects on estimation efficiency inherent in cluster sampling can be reduced if test forms are randomized over the units sampled from the same cluster (Johnson, 1992). A second difference encountered in assembling multiple parallel test forms is that in educational assessments, the use of test forms with different statistical features for different groups of students in the sample may be desirable. This option allows for the most optimal use of the tests as assessment instruments. Later in this paper, examples are given illustrating the use of this option.

This paper presents two different methods for the assembly of test forms for use in large-scale educational assessments. In the first method, the items are assigned directly from the pool to the individual test forms. The second method uses the current practice of the National Assessment of Educational Progress (NAEP), in which items are first organized as blocks, which are then assigned to the individual test forms following a balanced incomplete block design. The two methods are identical in that both use the technique of 0–1 linear programming (LP). The same technique was used earlier to solve such problems as matching a single test to a target information function, test assembly based on classical parameters, item matching, observed-score pre-equating, and item selection in constrained computerized adaptive testing. Some relevant references include Adema (1990, 1992), Adema, Boekkooi-Timminga, and van der Linden (1991),

> *. . . educational and psychological testing provides many cases in which the focal point is the assembly of sets of multiple test forms rather than a single form.*

Adema and van der Linden (1989), Amstrong and Jones (1992), Amstrong, Jones, and Wu (1992), Boekkooi-Timminga (1987, 1990), Theunissen (1985, 1986), van der Linden (1994, 1997, in press), van der Linden and Boekkooi-Timminga (1988, 1989), van der Linden and Luecht (1996, in press), and van der Linden and Reese (1998).

This paper assumes that information on background variables explaining the achievements of students in the assessment is known prior to the assembly of the tests. These variables can be used to define the various strata and clusters involved in the sampling design or special variables measured in the assessment. It is also assumed that the strata and clusters can be grouped according to their positions on these variables and that information from previous assessments can be used to derive prior ability distributions of these groups. Finally, it is likewise assumed that several forms are assembled for administration with groups with the same prior distribution. This assumption permits spiraling of test forms within groups to reduce the cluster effects. These assumptions underlie the sampling scheme presented in table 2.1. The term "cluster" in the table is used in the remainder of this paper as a generic term to denote a set of clusters of strata grouped on background variables.

The models also assume that the test forms are assembled from a pool of pretested items and that the pretest samples have been large enough to yield accurate estimates of such quantities as the item response theory (IRT) parameters of the items, their optimal response times, or other item attributes of interest. However, the methods will also work, albeit with less accuracy, for tentative estimates of these quantities.

The LP technique is used to assemble the test forms so that the background properties of the clusters of students are matched optimally with the properties of the pretested items. These items are subject to the various constraints that have to be imposed on the contents of the tests. Basically, the technique consists of the following steps:

1. Defining the decision variables that indicate whether an item should be assigned to a test form.
2. Using the variables to formulate a set of linear (in)equalities representing the constraints on the values of the variables that must be imposed.
3. Using the variables to formulate an objective function to optimize the tests.
4. Applying an algorithm or heuristic to calculate a feasible (i.e., admissible) set of values for the variables with an optimal value for the objective function.

It is not unusual for test assembly problems to have hundreds of constraints. The combinatorial complexity involved in assembling multiple test forms

Table 2.1. Population structure assumed

Cluster 1	Cluster 2	Cluster 3	Et cetera
Form A	Form D	Form F	
Form B	Form E	Form G	
Form C		Form H	
	Form I		

with such large numbers of constraints is already enough to motivate the application of 0–1 LP. Assembling several forms by hand, particularly if some of the constraints are tight, may take several days, whereas an appropriately implemented LP problem can be solved to a high degree of precision in minutes. If this complexity were the only motivation to use LP and the interest was exclusively in finding a feasible solution to the problem, the choice of objective function would be arbitrary and any convenient function would do. However, a large set of meaningful objective functions suggests itself for application in test assembly for educational assessments, including the following options:

1. If the interest is not only in estimating properties of the distributions of certain populations but also in reporting individual scores to schools, it may be helpful to maximize the efficiency of the individual ability estimators. The standard IRT approach in this case is to minimize the distance between the information functions of the test forms to targets for each cluster of students in the sample. The background information on the clusters available can be used to derive meaningful targets. This choice of objective function does not necessarily lead to better estimators of the parameters describing the ability distribution in a population of students, but gives an additional opportunity for optimization to improve the statistical features of the estimators of the individual θs. However, improved estimation of the θs makes marginal analysis of group differences more robust against model misspecifications (Mislevy, Beaton, Kaplan, & Sheehan, 1992).
2. The validity of educational assessments is low if the test forms do not motivate students to produce their best answers. One way to increase students' motivation is to give them items with probabilities of success that are neither too low nor too high. An objective function incorporating this idea is the one that minimizes the distance between the probabilities of success on the items for typical ability values of the clusters and target values for them.
3. Students vary in the time they need to produce a correct response to an item. At the same time, items differ in the time they require from a student. If individual differences in response time are considered a nuisance variable in the assessment, it may make sense to use an objective function that maximizes the match between the items and the students in the various clusters.

The translation of each objective into a linear objective function is demonstrated in this paper.

. . . an appropriately implemented LP problem can be solved to a high degree of precision in minutes.

If one or more of the clusters in table 2.1 (e.g., certain geographic areas) contained ability distributions to be estimated, a legitimate goal for each cluster would be minimization of a suitable function on the covariance matrix of the Marginal Maximum Likelihood (MML) estimators of the parameters characterizing its distribution. However, for the mainstream IRT models, such functions appear to be nonlinear in the items. Although in another multiparameter IRT problem an appropriate linearization of the objective function appears to be possible (van der Linden, 1996), no attempts have been made to deal with the current case.

The rest of this paper is organized as follows. First, a standard problem of optimal test assembly in IRT is formalized as an instance of 0–1 LP. This case is used to show how the first objective above can be given the shape of a linear objective function. In addition, it explains the different types of constraints that can be met in test assembly. The first model for the assembly of multiple forms for educational assessments is then given. In this model, the second objective is illustrated. The same technique of 0–1 LP is finally applied to optimize the assignment of blocks of items in a balanced incomplete block design. In this application, the last objective is shown.

0–1 Linear Programming Models for Test Assembly

It is assumed that the items in the pool are represented by decision variables x_i, i=1,..., I denoting whether (x_i=1) representing item i included in the test, or (x_i=0) representing the item i not included in the test. These variables model an objective function that minimizes the distances between the test information function and a target function over a series of values θ_k, k=1,..., K, in which $T(\theta_k)$ is used to denote the target values at these points. The values of the information function of item i at these points are represented by $I_i(\theta_k)$. In addition, examples from the following four categories of possible constraints are taken:

1. Constraints needed to fix the length of the test or the length of possible sections at prespecified numbers of items.
2. Constraints needed to model test specifications that deal with categorical item attributes. Examples of categorical item attributes are item content and format, the presence or absence of graphics, and the cognitive level of the item. The distinctive feature of categorical attributes is that each introduces a partition of the item pool with different classes of items associated with different levels of the attribute. The constraints in this category typically specify required distributions of items over the partitions.
3. Constraints needed to model test specifications that deal with quantitative item attributes. These attributes are parameters or coefficients with numerical values, such as item p-values, word counts, and (expected) response times. The constraints in this category usually require sums or averages of the values of these attributes to be in certain intervals.
4. Constraints needed to deal with possible dependencies of the test items in the pool. For example, certain items may have to be administered as a set related to the same text passage, whereas others are not allowed to figure in the same form because they have clues to one another.

For convenience, only one categorical item attribute (e.g., cognitive level) is used, with levels h=1,..., H, each corresponding to a different subset of items in the pool, C_h. For each subset, the number of items in the test has to be between $n_h^{(l)}$ and $n_h^{(u)}$. Likewise, one quantitative attribute is used, which is chosen to be the length of the items in the pool measured in numbers of lines, l_i.

The total number of lines of text in the test must not exceed the amount of space available, $l^{(u)}$. Finally, as an example of a dependency of the test items in the pool, it is assumed that items 83, 84, and 85 are not allowed in the same test form.

As can be seen, all expressions in the model are linear in the variables. Hence, models of this type can be solved for optimal values of their variables using a standard software package for LP. A choice of algorithms and heuristics is also offered in the test assembly package ConTEST (Timminga, van der Linden, & Schweizer, 1996). If the model has a special structure, efficient implementation of some algorithms may be possible (for an example, see Amstrong & Jones, 1992).

The model illustrates the use of the first objective for test assembly in assessments discussed above—namely, minimization of the distances between the information function of the test and a target for it. Although the distances at points θ_k are minimized from above, an approach in which the distances are minimized from below or from both sides is also possible.

The model runs as follows:

$$\text{minimize} \sum_{k=1}^{K} \Big[\sum_{i=1}^{I} I_i(\theta_k)x_i - T(\theta_k)\Big] \qquad \text{(target information function)} \qquad (1)$$

subject to

$$\sum_{i=1}^{I} I_i(\theta_k)x_i - T(\theta_k) \geq 0, \; k=1,\ldots, K \qquad \text{(positive differences)} \qquad (2)$$

$$\sum_{i=1}^{I} x_i = n, \qquad \text{(test length)} \qquad (3)$$

$$\sum_{i \in C_h} x_i \leq n_h^{(u)}, \quad h=1,\ldots, H, \qquad \text{(cognitive levels)} \qquad (4)$$

$$\sum_{i \in C_h} x_i \geq n_h^{(l)}, \quad h=1,\ldots, H, \qquad \text{(cognitive levels)} \qquad (5)$$

$$\sum_{i=1}^{I} l_i x_i \leq l^{(u)} \qquad \text{(number of lines)} \qquad (6)$$

$$x_{83} + x_{84} + x_{85} \leq 1 \qquad \text{(mutually exclusive items)} \qquad (7)$$

$$x_i + 0,1, \quad I=1,\ldots, I. \qquad \text{(definition of } x_i\text{)} \qquad (8)$$

Earlier Approaches to Multiple-Form Assembly

The first approach to the problem of assembling multiple test forms is assembling the forms in a sequential fashion, each time removing those items already selected from the pool and updating the model to fit the next problem. Versions of this approach are followed in many testing programs. However, the method has two serious disadvantages. Suppose, for example, that the problem is one of assembling a set of parallel test forms. If these forms are assembled one after the other, the value of the objective function for the solution to the model is likely to deteriorate with each succeeding form because the items with the best values for their attributes tend to be selected first. As a consequence, the forms would not be parallel. The second disadvantage is the possibility of unnecessary infeasibility of the problem at a later stage in the assembly process. This phenomenon is illustrated in table 2.2, which shows the levels of only a few of the items in a larger pool of two of the attributes.

For simplicity, these attributes are assumed to represent features that the items either have or do not have (e.g., use of graphics in the stem). Suppose that two test forms have to be assembled so that each form must have at least two items with attribute 1 and one item with attribute 2. In a sequential procedure, the selection algorithm might pick both item 2 and item 3 for the first test because they have large contributions to the target. However, as a consequence of this choice, a second form satisfying the same set of constraints is no longer possible. In a simultaneous approach, a sound algorithm would always assign item 2 to one test form and item 3 to the other, thus preventing this infeasibility. (In fact, it is the presence of such attribute structures, which often are not immediately obvious, that makes manual assembly of multiple test forms a notorious process in which, once a feasible solution is found, test assemblers may feel inclined to stop out of relief rather than the certainty that an *optimal* feasible solution has been found.)

Both disadvantages were already noted in Boekkooi-Timminga (1990). Her solution to the problem of assembling multiple parallel forms was to remodel the problem using different decision variables. If the individual forms are denoted by $f=1,\ldots,F$, the new decision variables, x_{if}, are defined such that $x_{if}=1$ indicates that item i is assigned to test form f and $x_{if}=0$ otherwise. Hence, each item is assigned directly to a test form and all assignments take place simultaneously. For the model in equations (1)–(7) the result would be as follows:

Table 2.2. Example of unnecessary infeasibility in sequential test assembly

	Item 1	Item 2	Item 3	Item 4	Item 5
Attribute 1	x		x	x	x
2		x	x		
Contribution to Tar	0.35	0.71	0.84	0.29	0.45

x indicates that the item has the attribute.

$$\text{minimize} \sum_{f=1}^{F} \sum_{k=1}^{K} [\sum_{i=1}^{I} I_i(\theta_k)x_{if} - T(\theta_k)] \qquad \text{(target information function)} \qquad (9)$$

subject to

$$\sum_{i=1}^{I} I_i(\theta_k)x_{if} - T(\theta_k) \geq 0, \quad k=1,\ldots, F, \; f=1,\ldots, F, \qquad \text{(positive differences)} \qquad (10)$$

$$\sum_{i=1}^{I} x_{if} = n, \quad f=1,\ldots, F, \qquad \text{(test length)} \qquad (11)$$

$$\sum_{i \in C_h} x_{if} \geq n_h^{(u)}, \quad h=1,\ldots, H, \; f=1,\ldots, F, \qquad \text{(cognitive levels)} \qquad (12)$$

$$\sum_{i \in C_h} x_{if} \geq n_h^{(l)}, \quad h=1,\ldots, H, \; f=1,\ldots, F, \qquad \text{(cognitive levels)} \qquad (13)$$

$$\sum_{i=1}^{I} l_i x_{if} \leq l^{(u)}, \quad f=1,\ldots, F, \qquad \text{(number of lines)} \qquad (14)$$

$$\sum_{f=1}^{F} x_{if} \leq 1, \quad I=1,\ldots, I, \qquad \text{(no overlap)} \qquad (15)$$

$$x_{83} + x_{84} + x_{85} \leq 1, \qquad \text{(mutually exclusive items)} \qquad (16)$$

$$x_{if} = 0, 1, \; i=1,\ldots, I, \; f=1,\ldots, F. \qquad \text{(definition of } x_{if}\text{).} \qquad (17)$$

Observe that equation (15) has been added to prevent each item from being assigned to more than two forms. The total number of constraints has also increased because all constraints in equations (9)–(14) are in force F times and equation (15) entails I new constraints. What is more important, however, is the fact that the number of variables has increased by a factor F. For this reason models of this type, although powerful for smaller problems, quickly result in memory management problems or prohibitively large computation times for more complicated problems. Therefore, the method presented in the next section is helpful.

Large-Scale Educational Assessments: Method 1

As already outlined, a basic problem with a sequential approach to assembling multiple forms is an unbalanced assignment of items to forms. Nevertheless, the approach does have the advantage of producing the smallest number of decision variables and constraints. The simultaneous approach discussed in the previous section deftly solves the problem of imbalance, but its price is a larger number of variables and constraints. The method in this paper, a version of which was already proposed by Adema (1990) for the assembly of a weakly parallel test, provides for the balancing of test content, while considerably minimizing the increase in the number of variables and constraints.

Basically, the method reduces any multiple-form assembly problem to a series of computationally less intensive, two-form problems. At each step, one form is assembled according to the true specifications. The other form is a dummy assembled according to adapted specifications; its only task is to balance the contents of the current form with later forms. As soon as both forms have been assembled, the items selected for the dummy are returned to the pool, and the process is repeated.

To present the model, two different sets of decision variables are used—one set of variables x_i, $i=1,\ldots, I$, to denote whether ($x_i=1$) or ($x_i=0$) item i will be assigned to the form assembled and another set of variables z_i, $i=1,\ldots, I$, for the same decision with respect to a dummy form. The objective is now the minimization of the distances between target values for the probabilities of success on the items and their likely values for the various clusters. To implement the objective, θ_f^* is a value typical of the abilities of the students in the cluster for which test form f is assembled. The values is assumed to be derived from background information on the students. In addition, the target value for these students is denoted as τ_f. The model for assembling form f plus its associated dummy is:

$$\text{minimize } y \qquad \text{(objective function)} \qquad (18)$$

subject to

$$P_i(+|\theta_f^*)x_i - \tau_f \leq y, \quad I=1,\ldots, I, \qquad \text{(target for form f)} \qquad (19)$$

$$P_i(+|\theta_f^*)x_i - \tau_f \geq -y, \quad i=1,\ldots, I, \qquad \text{(target for form f)} \qquad (20)$$

$$\sum_{g=f+1}^{F} [P_i(+|\theta_g^*)z_i - \tau_g] \leq [\sum_{g=f+1}^{F} n_g]y, \quad I=1,\ldots, I, \qquad \text{(target for dummy)} \qquad (21)$$

$$\sum_{g=f+1}^{F} [P_i(+|\theta_g^*)z_i - \tau_g] \geq -[\sum_{g=f+1}^{F} n_g]y, \quad I=1,\ldots, I, \qquad \text{(target for dummy)} \qquad (22)$$

$$\sum_{i=1}^{I} x_i = n_f, \qquad \text{(length of form f)} \qquad (23)$$

$$\sum_{i=1}^{I} z_i = \sum_{g=f+1}^{F} n_g, \qquad \text{(length of dummy)} \qquad (24)$$

$$\sum_{i \in C_h} x_i \leq n_{hf}^{(u)}, \quad h=1,\ldots, H, \qquad \text{(cognitive levels)} \qquad (25)$$

$$\sum_{i \in C_h} x_i \geq n_{hf}^{(1)}, \quad h=1,\ldots, H, \qquad \text{(cognitive levels)} \qquad (26)$$

$$\sum_{i \in C_h} z_i \leq \sum_{g=f+1}^{F} n_{hg}^{(u)}, \quad h=1,\ldots,H, \qquad \text{(cognitive levels)} \tag{27}$$

$$\sum_{i \in C_h} z_i \geq \sum_{g=f+1}^{F} n_{hg}^{(l)}, \quad h=1,\ldots,H, \qquad \text{(cognitive levels)} \tag{28}$$

$$\sum_{i=1}^{I} l_i x_i \leq l^{(u)}, \qquad \text{(number of lines)} \tag{29}$$

$$\sum_{i=1}^{I} r_i z_i \leq \sum_{g=f+1}^{F} r_g^{(u)}, \qquad \text{(number of lines)} \tag{30}$$

$$x_i + z_i \leq 1, \quad I=1,\ldots,I, \qquad \text{(no overlap)} \tag{31}$$

$$x_{83} + x_{84} + x_{85} \leq 1, \qquad \text{(mutually exclusive items)} \tag{32}$$

$$z_{83} + z_{84} + z_{85} \leq 1, \qquad \text{(mutually exclusive items)} \tag{33}$$

$$x_i \in \{0,1\} \quad i=1,\ldots,I. \qquad \text{(definition of } x_i\text{)} \tag{34}$$

$$z_i \in \{0,1\} \quad i=1,\ldots,I. \qquad \text{(definition of } z_i\text{).} \tag{35}$$

The constraints in equations (19) and (20) require the distances between the probabilities of success on the items and their target values to be in the interval $(-y,y)$. The same is done for the dummy test in equations (21) and (22), adapting for differences in test length. The size of the interval is minimized in equation (16). The general shape of this objective function is explained below. In equation (23) the length of form f is set equal to n_f items, whereas in equation (24) the length of the dummy is set equal to the sum of the lengths of all remaining forms. The constraints related to the various cognitive levels in equations (25)–(28) as well as to the total number of lines available for printing the test in equations (29) and (30) are adapted accordingly. The constraint needed to prevent the overlap of items between the test forms now has to be formulated as in equation (31). Finally, the constraints necessary to deal with dependencies of the items must be repeated for the dummy test in equation (33).

Note that the coefficients in the constraints have been made form-dependent to allow for differences in specifications between the forms. Apart from the change of variables, the main modifications in the constraints for the dummy test are on the right-hand side coefficients; these have been made larger to enforce adequate balancing of test contents between forms.

Objective Function

The objective function in the above problem along with its definition in equations (19)–(22) is of the minimax type—that is, for all items, a common upper bound to the distances between the probabilities of success and the target values is defined. This bound is next minimized. Application of the minimax principle is a convenient way to unify different objectives into a single objective function. Multiple-form test assembly problems always involve a distinct objective for each form. In this example, a different objective is involved for each individual item. Although attractive by itself, the use of objectives at item level is the only exception in which the suggestion contained in the following section is not expected to work satisfactorily.

Multiple-form test assembly problems always involve a distinct objective for each form.

Relaxed Decision Variables

Generally, if the problem is still too large to be solved in realistic time, an effective reduction of the computational complexity involved in 0–1 LP can be realized by relaxing the decision variables for the dummy test, that is, by replacing equation (35) by:

$$z_i \in [0,1] \quad i=1,\ldots, I. \tag{36}$$

This measure may result in a slightly less effective form of content balancing among the various test forms, but since the number of 0–1 variables is halved, the effect on the branch-and-bound step generally used in the algorithms and heuristics in this domain should be dramatic. Since some of the variables are now reals, the problem becomes the occurrence of mixed integer linear programming (MILP). However, as noted, this approach does not work satisfactorily if the objective addresses individual item attributes. In this example, an attempt is made to match the success probabilities and the targets by having the algorithm assign "partial items" to the dummy form if their actual probabilities of success at θ_f^* are too large.

Large-Scale Educational Assessments: Method 2

Another possible method of test assembly in large-scale assessments is derived from the two-stage method currently used by NAEP. In this method, the items in the pool are first assigned to a set of blocks and then the blocks are assigned to test forms (called "booklets" in NAEP). The assignment in the second stage follows a pattern known as a balanced incomplete block (BIB) design (Johnson, 1992). In this section, a 0–1 LP model for the assignment of blocks to booklets is formulated. The potential contribution of this model is not so much the possibility of automation but that it allows for better optimization of the design with respect to an objective function and the involvement of various other constraints in the assignment of blocks to booklets than those related to the parameters of the BIB design.

The following notation is needed to present the model. The blocks in the pool are represented by indices $i=1,\ldots, N$. To represent pairs of blocks a second index j with the same range of possible values is used. Booklets are denoted by $b=1,\ldots, B$. Decision variables x_{ib} are used to determine whether ($x_{ib}=1$) or not ($x_{ib}=0$) block i is assigned to booklet b. Likewise, variables z_{ijb} are used to assign pair (i,j) to booklet b. Special constraints will be formulated to keep the values of these two categories of variables consistent. The distribution of blocks across booklets is described by the following parameters:

- c_1 number of blocks per booklet;
- c_2 number of booklets per block;
- c_3 minimum number of booklets per pair of blocks.

To illustrate the possibility of controlling the contents of the booklets beyond the values of these parameters, three different kinds of additional constraints are introduced. First, it is assumed that the blocks are classified by content. Content is represented by a categorical attribute $c=1,\ldots, C$, where V_c is now defined as the subset of blocks in the pool belonging to content category c and n_c is the number of blocks to be selected from the pool. Second, it is assumed that the length of the booklets must be controlled. The number of lines of text in block i is denoted by a quantitative attribute l_i, whereas the total number of lines available for booklet b is $l_b^{(u)}$. Third, it is assumed that some blocks are "enemies" in the sense that they cannot be assigned to the same booklet. These sets of enemies are denoted by V_e, $e=1,\ldots, E$.

Finally, the model illustrates the use of an objective function based on response times needed for the items in the test. The variable r_{ib} can be the response time needed by the students in the cluster for which booklet b is assembled. An ideal definition of this parameter would be a certain percentile below the distribution of the actual response times in the cluster (e.g., 90th percentile). However, in practice it may be hard to estimate this parameter satisfactorily. A more practical choice, therefore, is to make an educated guess regarding the typical response time needed based on earlier experiences with the same type of students responding to the type of questions dominant in the booklet. The goal for the total response time needed for booklet b is T_b.

> *... the model illustrates the use of an objective function based on response times needed for the items in the test.*

The model is as follows:

$$\text{minimize } y \qquad \text{(objective function)} \tag{37}$$

subject to

$$\sum_{i=1}^{N} r_{ib}x_{ib} - T_b \leq y, \quad b=1,\ldots, B, \qquad \text{(ideal response time)} \tag{38}$$

$$\sum_{i=1}^{N} r_{ib}x_{ib} - T_b \geq -y, \quad b=1,\ldots, B, \qquad \text{(ideal response time)} \tag{39}$$

$$\sum_{i=1}^{N} x_{ib} = c_1 \quad b=1,\ldots, B, \qquad \text{(number of blocks per booklet)} \tag{40}$$

$$\sum_{i=1}^{N} x_{ib} = c_2, \quad I=1,\ldots, N, \qquad \text{(number of booklets per block)} \tag{41}$$

$$\sum_{b=1}^{B} z_{ijb} \geq c_3, \quad i<j=1,\ldots, N, \qquad \text{(number of booklets per pair)} \tag{42}$$

$$x_{ib} + x_{jb} \geq 2z_{ijb}, \quad i<j=1,\ldots, N,\ b=1,\ldots, B, \qquad \text{(consistent assignment)} \tag{43}$$

$$\sum_{b=1}^{B} \sum_{i \in V_c} x_{ib} \geq n_c, \quad c=1,\ldots, C, \qquad \text{(content)} \tag{44}$$

$$\sum_{i=1}^{N} l_i x_{ib} \leq l_b^{(u)}, \quad b=1,\ldots, B, \qquad \text{(length of booklet)} \tag{45}$$

$$\sum\sum_{(i<j) \in V_e} z_{ijb} \leq 1, \quad e=1,\ldots, E,\ b=1,\ldots, B, \qquad \text{(enemies)} \tag{46}$$

$$x_{ib} \in \{0,1\}, \quad i=1,\ldots, N,\ b=1,\ldots, B, \qquad \text{(definition of } x_{ib}\text{)} \tag{47}$$

$$z_{ijb} \in \{0,1\}, \quad i<j=1,\ldots, N,\ b=1,\ldots, B. \qquad \text{(definition of } z_{ijb}\text{).} \tag{48}$$

In equations (37)–(39) the minimax principle is applied again, this time to optimize the total response time needed for the booklets. The constraints in equations (40) and (41) define the size of the booklet by the numbers of blocks and the number of times a block is assigned to a booklet, respectively, whereas equation (42) sets the minimum number of booklets to which each possible pair is assigned as c_3. The constraints in equation (43) stipulate that each time a pair of blocks is assigned ($z_{ijb}=1$), it also holds that the individual pairs in this block are also assigned ($x_{ib}=1$ and $x_{jb}=1$). Observe that the reverse implication is not needed. Due to the constraints in equation (44), at least n_c blocks from content category have been assigned to a booklet, while the constraints in equation (45) guarantee that the length of booklet b is not longer than $l_b^{(u)}$ lines. Finally, the constraints in equation (46) prevent assigning more than one block from each set of enemies.

Discussion

The decision regarding which of the two methods should be recommended for use in large-scale assessments depends on such parameters as the size of the item pool, the number of test forms, and the number of blocks. Generally, the first method has only one parameter determining the number of variables in the model, namely, the size of the item pool. The number of test forms determines the number of iterative applications of the method. In the first application, the number of variables is equal to 2I. In the next application, n items have been removed from the pool, and the number of variables is 2(I–n). The number of variables in the second method depends on both the number of blocks and the number of booklets. More precisely, the method involves NB variables x_{ib} and N(N–1)/2 variables z_{ijb}.

For either method, the number of constraints depends on the number of attributes the test assembler wants to control. In addition, both methods involve technical constraints to keep the values of the different kinds of variables consistent. Unfortunately the number of such constraints depends directly on the size of the item pool or the numbers of blocks and booklets. If the number of constraints becomes too large, overflow of computer memory may occur.

The algorithms and heuristics in ConTEST (Timminga, van der Linden, & Schweizer, 1996) have been able to solve problems with 2,000 to 3,000 0–1 variables and several hundreds of constraints. For method one, these numbers imply that if relaxation of the z_i variables is possible, one can deal with item pools of this size. For method two, if N=20 and B=6, the numbers of variables and equations are typically between 1,000 and 1,500, and problems of this size also seem manageable. If the problem gets too large for either method, it may be split into subproblems dealing with disjoint parts of the item pool and solved separately. In fact, this measure has already been practiced by NAEP as "focused BIB design."

Finally, it is observed that, in principle, the problem of assigning items to test forms can be further refined by assigning items directly to positions in test forms. This approach would enable the pretest positions of the items to be taken into account if they could have a possible effect on the values of the item parameters. If no effects of pretest positions are present, the approach can be followed to neutralize possible effects of the position of the items in the assessment test on the results, such as the likelihood of the item not being reached or the student becoming tired, by systematically varying their positions across test forms. However, since the decision variables have to be indexed with respect to three different factors—items, forms, and positions—the increase in the number of variables needed can be kept within reasonable limits only if the possible positions of the items are categorized in a few larger classes (e.g., beginning, middle, and end of the test forms).

. . . the number of constraints depends on the number of attributes the test assembler wants to control.

Author's Note

Correspondence concerning this paper should be addressed to W. J. van der Linden, Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Electronic mail may be sent to vanderlinden@edte.utwente.nl.

References

Adema, J. J. (1990). The construction of customized two-staged tests. *Journal of Educational Measurement, 27,* 241–253.

Adema, J. J. (1992). Methods and models for the construction of weakly parallel tests. *Applied Psychological Measurement, 16,* 53–63.

Adema, J. J., & van der Linden, W. J. (1989). Algorithms for computerized test construction using classical item parameters. *Journal of Educational Statistics, 14,* 297–290.

Adema, J. J., Boekkooi-Timminga, E., & van der Linden, W.J. (1991). Achievement test construction using 0–1 linear programming. *European Journal of Operations Research, 55,* 103–111.

Amstrong, R. D., & Jones, D. H. (1992). Polynomial algorithms for item matching. *Applied Psychological Measurement, 16,* 365–373.

Armstrong, R. D., Jones, D. H., & Wu, I.L. (1992). An automated test development of parallel tests. *Psychometrika, 57,* 271–288.

Boekkooi-Timminga, E. (1987). Simultaneous test construction by zero-one programming. *Methodika, 1,* 101–112.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics, 15,* 129–145.

Johnson, E. G. (1992). The design of the national assessment of educational progress. *Journal of Educational Measurement, 29,* 95–110.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 131–161.

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika, 50,* 411–420.

Theunissen, T. J. J. M. (1986). Optimization algorithms in test design. *Applied Psychological Measurement, 10,* 381–389.

Timminga, E., van der Linden, W. J., & Schweizer, D. A. (1996). *ConTEST [Computer program and manual].* Groningen, The Netherlands: iec ProGAMMA.

van der Linden, W. J. (1994). Optimum design in item response theory: Applications to test assembly and item calibration. *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 308–318). G. H. Fischer & D. Laming (Eds.). New York: Springer-Verlag.

van der Linden, W. J. (1996) Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement, 20,* 373–388.

van der Linden, W. J. (1997). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement, 20,* 373–388.

van der Linden, W. J. (Ed.) (in press). Optimal test assembly. Special issue of *Applied Psychological Measurement.*

van der Linden, W. J., & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subsets method. *Applied Psychological Measurement, 12,* 201–209.

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 17,* 237–247.

van der Linden, W. J., & Luecht, R. M. (1996). An optimization model for test assembly to match observed-score distributions. *Objective measurement: Theory into practice,* (Volume 3, pp. 405–418). G. Engelhard & M. Wilson (Eds.). Norwood, NJ: Ablex Publishing Company.

van der Linden, W. J., & Luecht, R. M. (in press). Observed-score equating as a test assembly problem. *Psychometrika.*

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22.*

SECTION 3

A Brief Introduction to Item Response Theory for Items Scored in More Than Two Categories

David Thissen Kathleen Billeaud Lori McLeod Lauren Nelson

University of North Carolina at Chapel Hill

August 1997

A Brief Introduction to Item Response Theory for Items Scored in More Than Two Categories[*]

Many contemporary tests include constructed-response (CR) items, for which the item scores are ordered through categorical ratings provided by judges. When the judges' ratings use only two categories, widely known item response theory (IRT) models may be used. However, in most cases, responses to extended CR items or performance exercises are relatively long, and their scoring rubrics specify several graded categories of performance. The use of IRT with data from these kinds of items requires generalized models to accommodate the larger number of responses.

Alternatively, the responses to individual items on modern tests may not be locally independent, as required by the computations that produce IRT scale scores. Many reasons exist for local dependence. Several items may be based on a common stimulus: for example, the questions following a passage on a reading comprehension test, logical reasoning questions following a vignette, and mathematics questions based on some common graphic or illustration. CR items may be divided into parts that appear to be items. For example, a mathematics problem may be followed by a second item that asks for an explanation of the answer, or an examinee may be asked to make a drawing and then provide some written commentary on his or her art. While the parts that comprise these items may be scored separately, the item scores are likely to be correlated due to immediate associations related to the common stimulus or common aspects of the responses. Combining the parts of these locally dependent items into a larger unit, called a testlet (Wainer & Kiely, 1987), permits the use of the valuable machinery of IRT for item analysis and test scoring. Testlets, by nature, are large items that produce scores in more than two categories. They may use the same extensions of IRT as those developed for large items rated by judges in several scoring categories.

Testlets, by nature, are large items that produce scores in more than two categories.

In some cases, CR items or testlets may make up the entire test; in other cases, multiple-choice items are also used. In either case, a total score is often required, combining the judged ratings or testlet category scores and the binary item scores on the multiple-choice items if any are present. Simple summed scores may not be very useful in the latter context because of the problems associated with the selection of relative weights for the different items and item types and, in any event, because CR items are often on forms of widely varying difficulty. If the collection of items is sufficiently well represented by a unidimensional IRT model, scale scores may be a viable scoring alternative.

---

[*] Excerpts from a draft to appear in D. Thissen & H. Wainer (Eds.), *Test Scoring*—Chapter 4. A close relationship exists among various contemporary methods for standard setting and item response theory (IRT) scale scores for tests such as the National Assessment of Educational Progress. These brief excerpts from the volume *Test Scoring* (in preparation) on topics that arise in scoring tests that combine multiple-choice and constructed-response items may be useful in the explication of both scale scores and standard-setting methods that are based on data from different item types.

If the collection of items is sufficiently well represented by a unidimensional IRT model, scale scores may be a viable scoring alternative

One of the great advances of IRT over traditional approaches to educational and psychological measurement is the facility with which IRT handles items that are scored in more than two categories. Indeed, in the transition from dichotomously scored items to polytomously scored items, the only changes in IRT are the trace line models themselves. In this paper, the application of item response models to data in which the items have multiple (that is, more than two) possible scores will be considered.

Scale Scores for Items with More Than Two Response Categories

Estimates of Proficiency Based on Response Patterns

This section is very brief because the principles underlying the computation of scale scores using any of the polytomous models are identical to those widely used with IRT models for binary responses. The joint likelihood for any response pattern, $\mathbf{u} = \{u_1, u_2, u_3, \ldots\}$ is

$$L = \prod_{i=1}^{\text{nitems}} T_i(u_i | \theta)\, \phi(\theta)$$

regardless of whether each u represents a dichotomous or polytomous response category. The latent variable (proficiency) is denoted θ, and $\phi(\theta)$ is the population distribution [assumed to be N(0,1) in all the examples]. The only new point to be raised here is that the trace lines, $T_i(u_i | \theta)$, describing the probability of a response in category u_i for item i as a function of θ, may take different functional forms for the different item types and response formats: It does not matter if the trace lines arise from the one-, two-, or three-parameter logistic or from the graded model or from the nominal model or any of its specializations. MAP[θ] and EAP[θ] may still be computed exactly as they are for models for binary data.

IRT Analysis: The North Carolina Test of Computer Skills—Keyboarding

To illustrate the ideas involved in using IRT models for graded responses, we consider data from a 10-item North Carolina Test of Computer Skills, keyboarding (KB) section. Using the item response data from the 3,104 students who completed the item tryout forms, we fitted the graded (GR) model, the Generalized Partial Credit (GPC) model, and the Partial Credit (PC) model, with the computer program Multilog

(Thissen, 1991), all with a Gaussian population distribution for θ. The item parameter estimates for the GR model are listed in table 3.1, and those for the GPC model are listed in table 3.2. For these data, the GR model fits better than the GPC model. Both models use the same number of parameters, and -2log-likelihood is 105 for the GR model and 119 for the GPC model. (In this case, smaller is better.) Because these models are not hierarchically nested, no straightforward way exists to associate a probability statement with the fact that the GR model provides a better fit. Nevertheless, that is the fact.[1] The PC model (constraining all the item discrimination parameters to be equal) is testable because it is hierarchically nested within the GPC model. The test of the null hypothesis that the slopes are equal is rejected: $G^2(2)=7$, $p = 0.03$.

The trace lines for the GR and GPC models are shown in figure 3.1. Note that the trace lines for the two models are very similar. The small differences in the shapes of the curves account for the difference between the goodness-of-fit of the two models, but has little consequence for scoring. Tables 3.3 and 3.4 show the values of EAP[θ] and standard deviation (*s.d.*)[θ] for the 27 response patterns, using the GR and GPC models, respectively. For the most part, scale scores for the same response pattern from either of the two models differ by less than 0.1 standard units. Exceptions are response patterns that involve some points for items 1 and 2, but 0 for item 3 (010, 110, and 220). Those relatively rare patterns have scale scores that are 0.1 to 0.2 standard units different for the GPC model than for the GR model.

Table 3.1. Graded (GR) model item parameters for the North Carolina Test of Computer Skills, keyboarding (KB) section

Item	Item Parameters		
	a	b_1	b_2
1. Typing	1.05	−1.40	−0.97
2. Spacing	1.59	−0.52	0.92
3. Length	0.93	−2.97	0.55

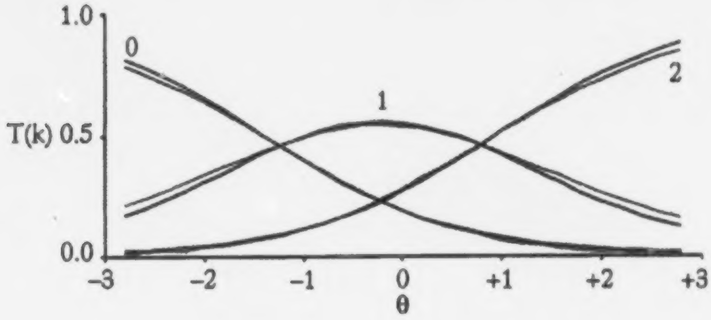
Table 3.2. Generalized Partial Credit (GPC) model item parameters for the North Carolina Test of Computer Skills, KB section

Item	Item Parameters					
	a_1	a_2	a_3	c_1	c_2	c_3
1. Typing	−0.84	0.0	0.84	−0.47	0.57	−0.10
2. Spacing	−1.19	0.0	1.19	0.04	0.42	−0.46
3. Length	−0.81	0.0	0.81	−1.40	0.90	0.50

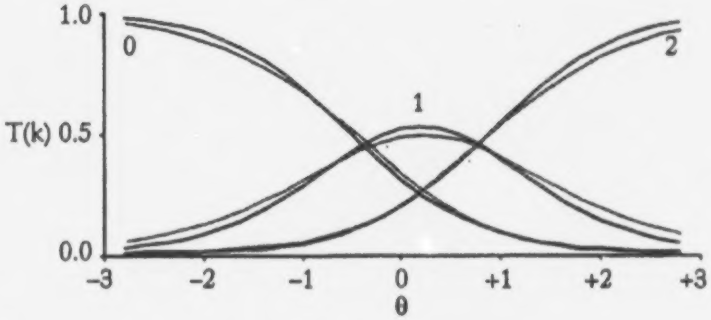
---

[1] In our experience, fitting hundreds of data sets over two decades, it has almost always been the case that the GR model fits rating data better than the GPC model. The difference is usually small, as it is in this case.

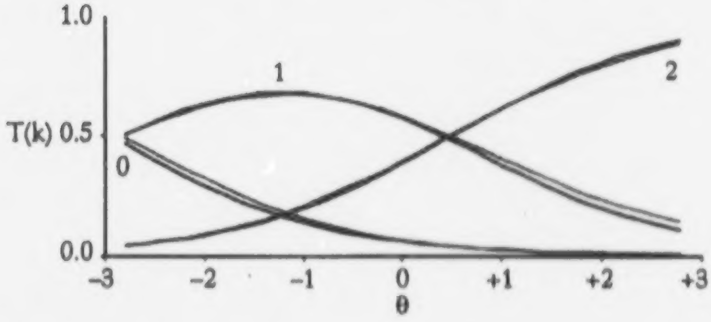
Figure 3.1. The trace lines for the GR (black) and GPC (gray) models for the North Carolina Test of Computer Skills, keyboarding items



"Typing"



"Spacing"



"Length"

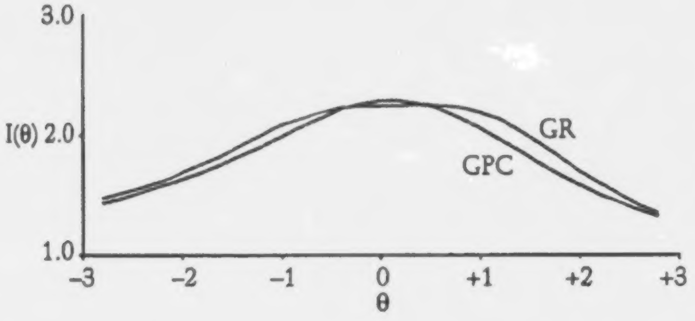
Table 3.3. EAP[θ] response-pattern scale scores and their corresponding *s.d.* and frequency, using the GR model for the North Carolina Test of Computer Skills, KB section

Response Pattern	Summed Score	EAP[θ]	*s.d.*[θ]	Frequency
000	0	-1.55	0.76	64
001	1	-1.12	0.72	236
100	1	-1.03	0.72	25
010	1	-0.74	0.70	31
002	2	-0.71	0.73	96
101	2	-0.68	0.69	337
200	2	-0.67	0.76	17
011	2	-0.43	0.67	142
110	2	-0.34	0.67	58
020	2	-0.30	0.80	8
201	3	-0.31	0.72	137
102	3	-0.28	0.70	138
111	3	-0.08	0.64	340
012	3	-0.04	0.67	57
210	3	0.03	0.69	27
021	3	0.05	0.75	37
120	3	0.15	0.74	11
202	4	0.15	0.74	76
211	4	0.27	0.66	167
112	4	0.27	0.64	212
121	4	0.42	0.70	138
022	4	0.54	0.75	36
220	4	0.65	0.76	14
212	5	0.66	0.67	126
122	5	0.85	0.70	196
221	5	0.87	0.71	97
222	6	1.34	0.73	281

Table 3.4. EAP[θ] response-pattern scale scores and their corresponding *s.d.* and frequency, using the GPC model for the North Carolina Test of Computer Skills, KB section

Response Pattern	Summed Score	EAP[θ]	*s.d.*[θ]	Frequency
000	0	-1.53	0.75	64
001	1	-1.09	0.72	236
100	1	-1.07	0.72	25
010	1	-0.89	0.71	31
002	2	-0.67	0.70	96
101	2	-0.66	0.70	337
200	2	-0.65	0.70	17
011	2	-0.49	0.69	142
110	2	-0.48	0.69	58
020	2	-0.31	0.69	8
102	3	-0.27	0.69	138
201	3	-0.26	0.69	137
012	3	-0.10	0.69	57
111	3	-0.09	0.68	340
210	3	-0.08	0.68	27
021	3	0.07	0.68	37
120	3	0.08	0.68	11
202	4	0.13	0.68	76
112	4	0.29	0.69	212
211	4	0.30	0.69	167
022	4	0.45	0.69	36
121	4	0.47	0.69	138
220	4	0.48	0.69	14
212	5	0.69	0.70	126
122	5	0.86	0.71	196
221	5	0.88	0.71	97
222	6	1.30	0.74	281

Figure 3.2. Test information curves for the North Carolina Test of Computer Skills, KB section, as computed using both the GR and the GPC models



Test information curves for this three-item test, as computed using both the GR and GPC models, are shown in figure 3.2. As one would expect given the similarity of the trace lines, the two information curves in figure 3.2 are very similar. The striking feature of the two curves is that they are both very flat, across a wide range of θ. Thus, IRT displays the primary advantage of multiple-category scoring: Each item provides information at two (in this case) or more (in general) levels of proficiency. By adjusting the definitions of the scoring categories, a test can be reasonably easy to construct that measures proficiency almost equally accurately over a wide range of proficiency. For most values of θ near the middle of the scale, the value of test information is about 2.0. Therefore, the standard errors of the scale scores are expected to be about $\sqrt{1/2} \cong 0.7$, as they are in tables 3.2 and 3.3.

Using IRT Scale Scores for Response Patterns to Score Tests Combining Multiple-Choice and Constructed-Response Sections: Wisconsin Third-Grade Reading Field Test

To illustrate the use of IRT scale scores associated with response patterns to score tests comprising both multiple-choice and CR sections, we use data from a field test of a form of Wisconsin's third-grade reading test. The data were obtained from 522 examinees. Here we use the responses to 16 multiple-choice items, as well as the responses to 4 CR items. All 20 items followed a single reading passage. The multiple-choice items were in a conventional four-alternative format; the CR items were open-ended (OE) questions that required a response on a few lines. The OE items were rated on a four-point scale (0–3). We simultaneously fitted the 3PL model to the multiple-choice items and the GR model to the OE items using Multilog (Thissen, 1991) and using a mild Bayesian prior distribution[2] for the guessing parameter of the 3PL model (*g*).

Figure 3.3 illustrates the computation of the IRT scale score for an individual examinee for this 20-item test. The examinee responded correctly to 12 of the 16 multiple-choice items, obtained scores of 2 on 2 of the OE items, and 3 on the other 2, for a total score (if one sums the "points") of 22. IRT response-pattern scoring ignores the summed score and instead multiplies the trace lines shown in figure 3.3. The top panel shows the trace lines for the examinee's responses to the multiple choice items (12 increasing curves and 4 decreasing curves, for the 12 correct and 4 incorrect item responses); the middle panel shows the trace lines associated with the OE responses. The 2 nonmonotonic curves are the GR trace lines for the 2s, and the 2 increasing trace lines are those for the 3s. The bottom panel of figure 3.3 is the product of the 20 curves in the other 2 panels [and the N(0,1) population distribution curve]. The mode of the curve in the lower panel is MAP[θ], which takes a value of –0.54 (with *s.e.* = 0.27)—that is, the IRT scale score in *z*-score units for this examinee.

IRT scale scores computed as in figure 3.3 for each examinee provide a solution to the "weighting problem" for tests such as this one that combine multiple-choice and OE items. Many would question the use of summed "points" to score a test such as this one, asking why one of the rated "points" for the OE responses should equal the value of a correct multiple-choice response. However, the IRT scale scoring process neatly finesses the issue: All of the item responses (open-ended and multiple-choice) are implicitly "weighted"; indeed, the effect of each item response on the examinee's score depends on the other item responses. Each response pattern is scored in a way that best uses the information about proficiency that the entire response pattern provides, assuming that the model summarizes the data accurately.

Each response pattern is scored in a way that best uses the information about proficiency that the entire response pattern provides . . .

IRT scale scores computed in this way may vary a good deal for examinees with the same summed score. Figure 3.4 shows a scatter diagram of the scale scores for the combined 3PL and GR model, plotted against the summed score computed, taking the number of OE rated "points" literally.[3] For some summed scores, the range of IRT scale scores is as much as a standard unit. A good deal of this range is attributable, in this case, to the differential treatment by the IRT model of the OE responses. As shown in figure 3.3, the slopes of the trace lines[4] for the OE responses are substantially less than the slopes of

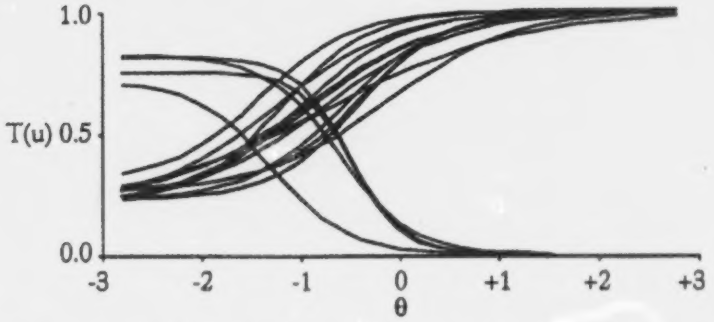
---

[2] The prior distribution used for the lower asymptote parameter for the four-alternative multiple-choice items was N(-1.1, 0.5) for logit(*g*). This distribution has a mode of 0.25 for *g*.

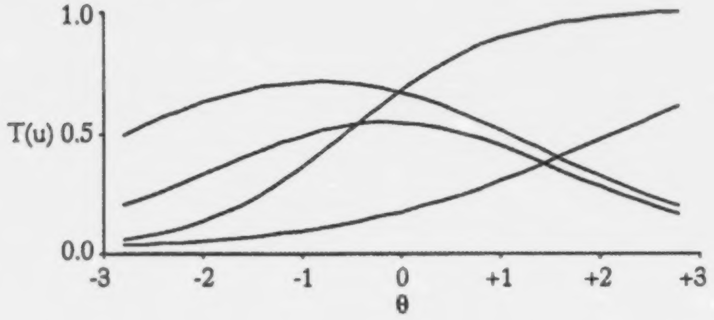
[3] Examinees with missing responses are omitted from figure 3.4 because it is not clear how to compute their summed scores in a way that is comparable to those of other examinees.

[4] Here, "slope" is taken generally to mean the rate of change of the response probability as a function of θ.

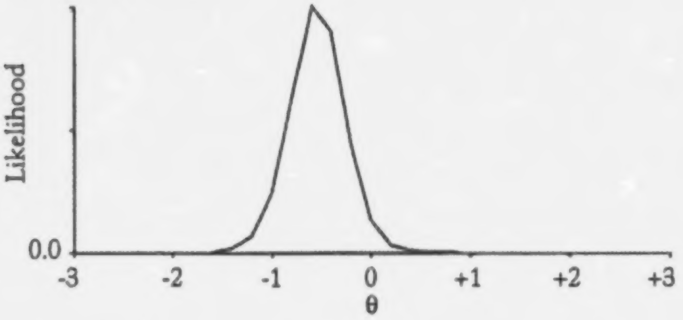
Figure 3.3. The computation of the scale score for an individual examinee on the Wisconsin third-grade reading test



The top panel shows the 3PL trace lines for this examinee's responses to the multiple-choice items



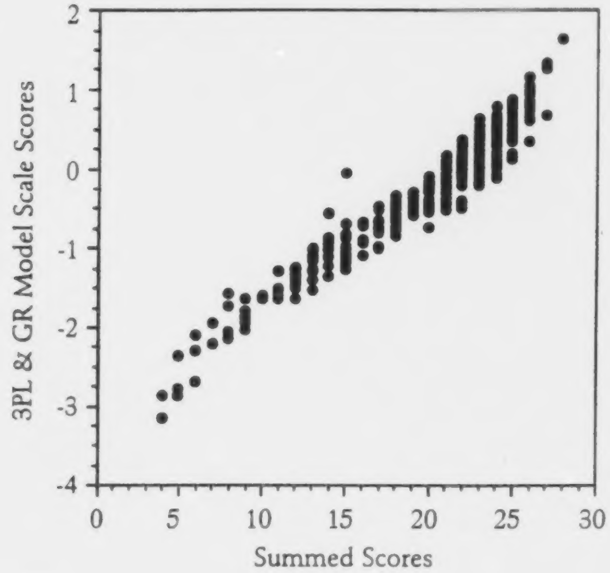
The middle panel shows the GR model trace lines associated with the open ended (OE) responses



The bottom panel shows the product of the 20 curves in the other two panels and the N (0,1) population distribution curve

the 3PL trace lines in the vicinity of this examinee's score. As a result, the 3PL responses "count" more in the score, and the OE responses "count" relatively less. For examinees other than the one shown in figure 3.3 who also obtained a summed score of 22, the IRT scale score is higher because they responded correctly to more of the highly discriminating multiple-choice items, even though they obtained fewer points for their OE responses.

Figure 3.4. Scatter diagram of the scale scores for the combined 3PL and GR model, plotted against the summed scores for the Wisconsin third-grade reading test



Assuming that the combined 3PL and GR model accurately represents the data, the IRT scale scores simultaneously provide more accurate estimates of each examinee's proficiency and avoid any need for explicit consideration of the relative "weights" of the different kinds of "points."

The Testlet Concept

The concept of the testlet was introduced in the literature by Wainer and Kiely (1987, p. 190): "A testlet is a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow." Wainer and Kiely proposed the use of testlets as the units of construction and analysis for computerized adaptive tests (CATs). However, the testlet concept is now viewed as a general-purpose solution to the problem of local dependence (LD) (Yen, 1993). If a pair or cluster of items exhibits LD with respect to the construct being measured by the test as a whole, that pair or cluster of items may be aggregated into a single unit—a testlet. The testlet then yields locally independent responses in the context of the other items in the measure. Testlets and individual items can then be included in an IRT model for item analysis and test scoring.

By definition, a testlet is a kind of (super) test item that yields more than two responses; furthermore, the relative ordering of those responses with respect to the construct being measured may or may not be known a priori. Although traditional approaches to item analysis and test assembly may be stymied by the presence of items with multiple, purely nominal responses, Bock's (1972) IRT model for responses in several nominal categories may be used to provide

By definition, a testlet is a kind of (super) test item that yields more than two responses. . . .

straightforward item analysis and test scoring. Analysis using the NO model can also be used to determine if the responses are, as a matter of empirical fact, ordered. If they are, then a constrained version of the NO model, like the GPC or PC model or Samejima's (1969) GR model, may be effectively used.

Our first illustration of the testlet idea (Thissen & Steinberg, 1988; Thissen, 1993) used data reported by Bergan and Stone (1985) involving the responses to four items measuring the numerical knowledge of a sample of preschool children and the nominal IRT model. This example represents what we now call response-pattern testlets, in which every pattern of responses to the items in the testlet becomes a response category for the testlet as a whole. An extensive theoretical treatment of processes that may be represented by fitting the NO model to response-pattern testlets has recently been provided by Hoskens and De Boeck (1997), who reanalyze the Bergan and Stone example and also contribute others.

. . . testlet-based IRT analysis yields a more accurate description of the reliability and scale score standard errors for such tests.

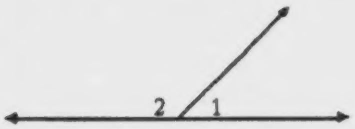
No loss of information is involved in the construction of response-pattern testlets; the data are merely redefined. However, given current technology, response-pattern testlets may include only a few items (two, three, or perhaps four) and only a few responses for each item because the number of response categories for the testlet is the product of the numbers of response categories for the items. That number cannot be larger than, say, 16 and still permit any reasonable amount of data to be used to calibrate the NO model.

When several items follow a common stimulus, it may be better to view the summed score on the test as the sum of the number-correct scores for the subsets of items associated with each of the common stimuli (often passages in reading comprehension tests). Both the nominal and GR models have been used for the analysis of passage-based tests (Thissen, Steinberg, & Mooney, 1989; Wainer, Sireci, & Thissen, 1991). To implement this idea, a test with 10 questions following each of four reading passages is treated as a four-testlet test, and the GR model or some version of the NO model is fitted to the 11 response categories that represent each possible number-correct score. In this case, some loss of information occurs relative to full response-pattern analysis of the test data. However, the loss of information is usually small and is more than compensated for since the testlet analysis is a proper analysis of locally independent responses, while the response-pattern analysis may be distorted by the LD induced by the passages. For this reason, testlet-based IRT analysis yields a more accurate description of the reliability and scale score standard errors for such tests (Sireci, Thissen, & Wainer, 1991).

The reason that pairs or clusters of items should be treated as testlets is sometimes obvious, as in the case of clustered tasks—for example, the pairs of questions on the preschool numerical knowledge test or the questions following a passage on a reading comprehension test. On the other hand, local dependence may also be an unexpected or even surprising empirical

Figure 3.5. Two items (14 and 15) from the 1991 North Carolina End-of-Course Geometry Test that exhibit substantial local dependence

14. Which term describes ∠1 and ∠2?



A vertical
B complementary
C adjacent
D congruent

15. If $m\angle A = 60$, what is the measurement of the supplement of ∠A?

A 120
B 90
C 40
D 30

phenomenon. Yen (1993) and her colleagues at CTB/McGraw-Hill have successfully used empirical procedures to detect LD on recently constructed performance-based educational assessments. They used a statistic, called Q_3, proposed by Yen (1984), to identify LD. However, Q_3 may exhibit unpredictable behavior under some circumstances (Chen & Thissen, 1997; Reese, 1995). For binary test items, we use the LD index described in detail by Chen and Thissen (1997). The LD index provides a straight- forward analysis of the residuals from the IRT model for each pair of items. If the items are locally independent, then the residuals from the fitted model for each pair of items are statistically independent, and the χ^2-distributed statistic is expected to be approximately one. If the items are locally dependent, the LD index is large. When substantial LD is detected—for example, by the LD index—or expected because the test was deliberately constructed with related items, we combine those items into testlets and use the same IRT machinery for test scoring that we use with any test that has items with several response categories.

Using the Nominal Model for Items Combined into Response-Pattern Testlets: An Example from the North Carolina End-of-Course Geometry Test

A pair of items that exhibit relatively extreme local dependence is shown in figure 3.5. These two items were numbers 14 and 15 on a 1991 test form of the North Carolina End-of-Course Geometry Test and physically adjacent, as shown in the figure. Such physical proximity often exacerbates LD when it exists. Using data from 2,739 examinees, we calibrated the 60-item test and computed the LD index for each pair of items. The value of the LD index for this pair of items was 180. For a statistic that is distributed approximately as a χ^2 with 1 *d.f.* in the null case, that is remarkable.

Many more examinees than predicted by the IRT model respond correctly to both items or incorrectly to both items. Examining the items, it is easy to see why. Indeed, several ways can be used to describe the probable reasons for the LD. A succinct description would be that the two items are both "vocabulary" items on a geometry test, and students whose teachers emphasized the memorization of vocabulary would do better on both. A somewhat more elaborate chain of reasoning would explain that item 14 could serve to "give away" the answer to item 15, even for students with limited knowledge: If we assume that among the three terms *adjacent, complementary,* and *supplementary,* the first is the easiest to remember, then we may assume that item 14 is easy to answer correctly (by selecting *adjacent*). However, the figure clearly shows that the angles in item 14 sum to 180°. That implies that *complementary,* as an incorrect distractor for item 14, cannot be the word that means "sums to 180°." When we turn to item 15, if we remember that supplementary is "one of those words" and means either "sums to 180°" or "sums to 90°," we eliminate alternatives B and C, and then (correctly) choose alternative A because *supplementary* has to be "sums to 180°" if (from item 14) *complementary* is not.

In any event, the empirical fact is that the responses to the two items are not locally independent. The solution proposed by Yen (1993) for this kind of LD is to combine the two items into a single testlet and conduct the IRT analysis again. This serves to eliminate the LD and keep the IRT model and its usefulness for scale scoring. (The alternatives are to keep the LD and eliminate the unidimensional IRT model or grossly complicate the model; neither of these ideas is attractive.) In this case, we rescored these two items as a single testlet with four response categories: 0 for response pattern {00}, 1 for response pattern {01}, 2 for response pattern {10}, and 3 for response pattern {11}. We then recalibrated the test using the 3PL model for the remaining 58 items and the NO model for the testlet.

Figures 3.6–3.8 illustrate response pattern items 14 and 15. The NO model trace lines for the four response patterns to items 14 and 15 are shown in figure 3.8. The trace lines for {00}, {01}, and {10} are all monotonically decreasing (and nearly proportional to one another) over the useful range of θ. Of course, the trace line for {11} is monotonically increasing. This differs from any pattern that can be obtained by combining two 3PL trace lines, for which the likelihoods of the four response patterns must be ordered, with {11} associated with higher values of θ, {01} and {10} associated with some intermediate values of θ, and {00} associated with the lowest. The 3PL trace lines for items 14 and 15 are shown in figure 3.6, and the products of those two curves (and their complements) are shown in figure 3.7. In an attempt to fit the data as accurately described by the NO in figure 3.8, the 3PL model has extremely high values of *g* (the lower asymptote) for both items. (They are 0.43 and 0.48, respectively.) Nevertheless, the 3PL model must imply that the likelihood for the response patterns {01} and {10}

Figure 3.6. 3PL trace lines for responses to items from the 1991 North Carolina End-of-Course Test in Geometry that exhibit substantial local dependence

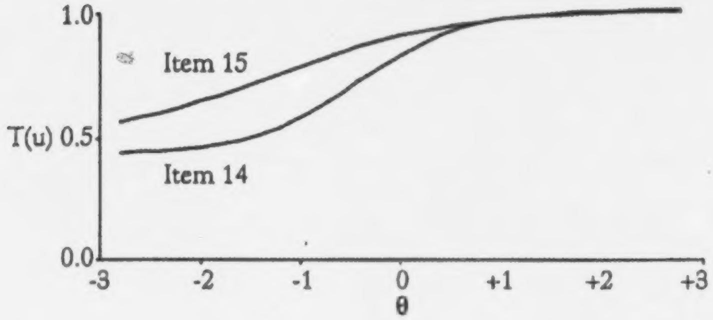


Figure 3.7. The likelihoods for the four possible response patterns for items 14 and 15, computed as products of the trace lines in Figure 3.6 and their complements

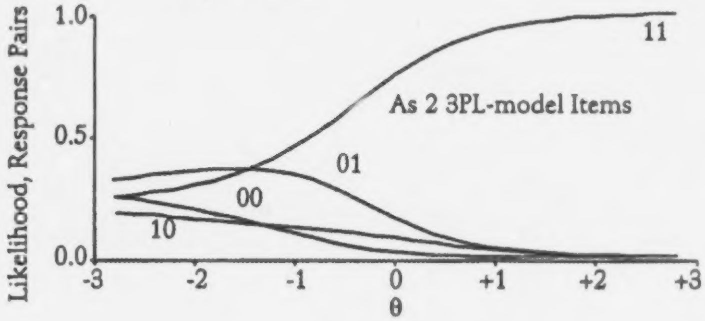
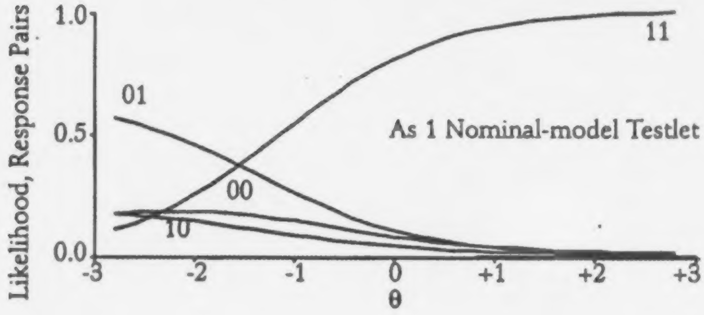


Figure 3.8. The nominal model trace lines for the four response patterns to items 14 and 15 of the 1991 North Carolina End-of-Course Test in Geometry



must be associated with relatively higher values of θ than {00}. This causes the misfit with the data that is detected by the LD index.

If the 3PL trace lines were used to compute IRT scale scores, the effect would be that examinees who responded correctly to either one of the two items would receive higher scale scores than those who responded correctly to neither. However, trace lines from the NO model for the testlet have different consequences for scale scoring: Effectively, the examinee "gets credit" (i.e., the scale score increases) for response pattern {11}, but the scale score tends to decrease slightly for any of the patterns that include an incorrect response. The testlet combination and subsequent NO model analysis have created a scoring rule for this pair of items that, to anthropomorphize, basically says, "This pair is a single item; you get one point if you respond correctly to both questions and zero points if you miss either." The IRT analysis indicates that using this scoring rule makes this pair of items a better indicator of proficiency in geometry than any that would be obtained treating the two items separately.

... trace lines from NO model for the testlet have different consequences for scale scoring.

We have described response-pattern testlet modeling and scoring as an a posteriori "fix" for observed LD, and the procedure is used this way. However, as Hoskens and De Boeck (1997) and others have pointed out, the availability of this kind of analysis makes it possible for the test constructor to plan or intend to construct item combinations as testlets. An example could be the mathematics item format that asks for the solution to a problem, and then follows up with what appears to be a second question, "Explain your answer." After the solution is scored (possibly as correct or incorrect) and the explanation is rated by judges following some rubric (perhaps on a three-point scale), all of the patterns of {solution score, explanation score} can be treated as the response alternatives for a single testlet, fitted with the NO model.[5]

Conclusion

IRT models for items with responses scored in more than two categories provide a useful way to compute scale scores for the otherwise difficult data that arise in the context of performance assessments. While the rating categories used by judges to provide the item-level scores for CR items are often arbitrary, the IRT models provide a mechanism to combine those ratings into scores on a useful scale. Weighting problems, once the province of guesswork or committees, are naturally handled in the process of the computation of IRT scale scores: each item is implicitly weighted according to its relation with the aspect proficiency that is common to all of the items (θ).

The computation of IRT scale scores requires that the "item" responses must be locally independent because that justifies the multiplication of the trace lines for those responses to compute the likelihood that is the basis of the scale scores. While this requirement may seem at first look restrictive in the context of performance assessment, we have seen that the testlet concept may be used to combine "items" (from the point of view of the examinee) that may be locally dependent into testlets in such a way that the testlet responses are locally independent. Then IRT models for responses in more than two categories, such as those we have discussed in this chapter, may be used to gain all the advantages of IRT—a well-defined scale with well-defined standard errors, form linking, and even adaptive testing.

---

[5] Some testing programs use ordered versions of the NO model, such as the GPC model, for this purpose. That may be effective, but we would recommend that the unconstrained NO model be fitted, or some other analysis performed, to check that the empirical order of the testlet categories corresponds to the order assumed in fitting an ordered item response model.

IRT scale scores may always be computed using the full response pattern, and if the model is well chosen for the data, one that yields the most statistically efficient scores. In some contexts, IRT scale scores may usefully be computed for each summed score. This is often more practical in large-scale testing programs but not as useful when the items being considered are of very different kinds.

References

Bergan, J. R., & Stone, C. A. (1985). Latent class models for knowledge domains. *Psychological Bulletin, 98,* 166–184.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37,* 29–51.

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265–289.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods, 2,* 261–277.

Reese, L. M. (1995). *The impact of local dependencies on some LSAT outcomes.* LSAC Research Report Series. Newtown, PA: Law School Admission Council.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Thissen, D. (1991). *Multilog user's guide—Version 6* [Computer program]. Chicago: Scientific Software, Inc.

Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. In N. Frederiksen, R. J. Mislevy & I. Bejar (Eds.). *Test theory for a new generation of tests* (pp. 79–97). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104,* 385–395.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26,* 247–260.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197–219.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–214.

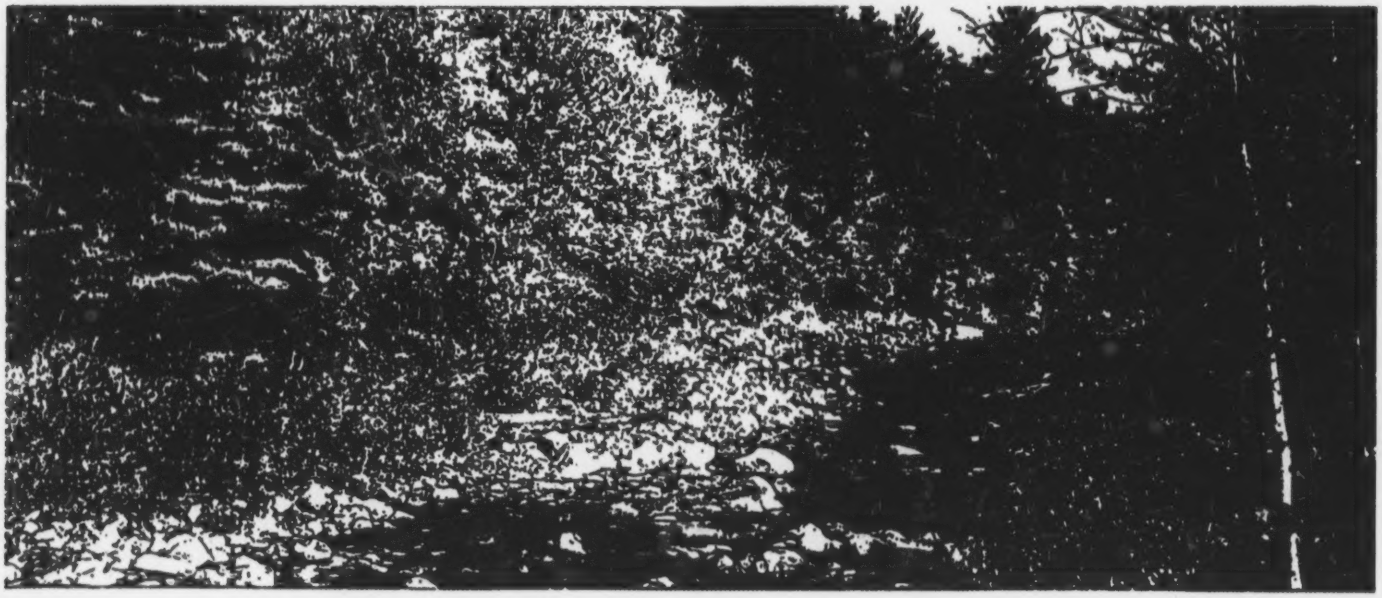
SECTION 4

Some Ideas about Item Response Theory Applied to Combinations of Multiple-Choice and Open-Ended Items: Scale Scores for Patterns of Summed Scores

Kathleen Billeaud   Kimberly Swygert   Lauren Nelson   David Thissen

University of North Carolina at Chapel Hill

August 1997



Some Ideas about Item Response Theory Applied to Combinations of Multiple-Choice and Open-Ended Items: Scale Scores for Patterns of Summed Scores*

Many contemporary tests include open-ended (OE) items, for which the item scores are ordered categorical ratings provided by judges, as well as multiple-choice items. If the collection of items is sufficiently well represented by a unidimensional item response theory (IRT) model, scale scores may be a viable plan for scoring such a test. Either EAP or MAP estimates of θ based on response patterns are traditional IRT answers to the scoring of such tests. However, in many contexts, response-pattern scoring carries nonpsychometric penalties, and an alternative solution is required. Weighted combinations of the summed scores are widely used, but no clearly superior solution exists to the problem of selecting the weights.

Scale Scores Based on Patterns of Summed Scores

A better solution to the problem of combining binary-scored, multiple-choice (MC) sections with items scored in multiple categories may involve a hybridization of summed-score and response-pattern computation of scaled scores. To do this, one must first jointly calibrate all the items (that is, one estimates the item parameters), using suitable IRT models for each item. Then one computes $L_x^{MC}(\theta)$, the likelihood for summed score x for the MC section, and $L_{x'}^{OE}(\theta)$, the likelihood for summed score x' for the OE section. For each combination of a given summed score x on the MC section with any summed score x' on the OE section, compute the product

$$L_{xx'}(\theta) = L_x^{MC}(\theta)\, L_{x'}^{OE}(\theta)\, \phi(\theta) \qquad (1)$$

The product in equation (1) is the likelihood for the response pattern defined as {score x on the MC section and score x' on the OE section}. Then we can compute the modeled probability of the response pattern of summed scores $\{x, x'\}$,

$$P_{xx'} = \int L_{xx'}(\theta)\, d\theta \qquad (2)$$

We may also compute the expected value of θ, given the response pattern of summed scores $\{x, x'\}$,

$$\mathrm{EAP}[\theta \mid x, x'] = \frac{\int \theta\, L_{xx'}(\theta)\, d(\theta)}{P_{xx'}} \qquad (3)$$

---

*Excerpts from a draft to appear in D. Thissen and H. Wainer (Eds.), *Test Scoring*.

and the corresponding standard deviation,

$$s.d.[\theta|x\,,\,x'] = \left[\frac{\int(\theta - EAP[\theta|x\,,\,x'])^2\,L_{xx'}(\theta)\,d(\theta)}{P_{xx'}}\right]^{1/2} \quad (4)$$

Equations (3) and (4) define two-way score translation tables that provide scaled scores and their standard errors for each such response pattern, in which the "pattern" refers to the ordered pair {score x on the MC section, score x' on the OE section}. This procedure offers many of the practical advantages of summed-scores, while preserving the differences in scale scores that may be associated with very different values of "points" on the MC and OE sections.

Using IRT Scale Scores for Patterns of Summed Scores to Score Tests Combining Multiple-Choice and Constructed-Response Sections: Wisconsin Third-Grade Reading Field Test

To illustrate the construction of scale scoring tables using equations (3) and (4), we use data from the Wisconsin third-grade reading test. We fitted the 16-item MC section and 4 OE items, each rated 0–3, with the 3PL and GR models (Birnbaum, 1968; Samejima, 1969) respectively, using the computer program Multilog (Thissen, 1991) and computed scale scores for the response patterns to these 20 items. Here, we use the item response models and item parameter estimates to compute the values of $EAP[\theta|x\,,\,x']$ and $s.d.[\theta|x\,,\,x']$, using equations (3) and (4).

Table 4.1 shows the values of $EAP[\theta|x\,,\,x']$ for the 221 combinations of x, the summed score on the MC section, and x', the summed score on the OE section. Tabulations such as those shown in table 4.1 can be used in score-translation systems; one enters the table with the summed scores on the two parts of the test and locates the scale score for that combination in the body of the table. A similar array of the values of $s.d.[\theta|x\,,\,x']$ is shown in table 4.2; these may be reported as the standard errors of the scores.

Table 4.1 shows some interesting features of IRT-based score combination, as opposed to the more commonly used simple weighted combinations of summed scores. Reading across each row or down each column of table 4.1, it should be noted that the "effect" of a "point" on either the OE section or the MC section depends on its context. In a simple weighted combination of the scores, obtaining 12 points instead of 11 on the OE portion would have the same "effect" on the score as obtaining 1 point instead of 0. This is not true for the IRT system: The likelihood-based system "considers" (to anthropomorphize) the two scores and their consistency pieces of evidence. Where the two pieces of evidence essentially agree

Table 4.1. The values of *EAP*[θ|*x* , *x'*] for combinations of multiple-choice (MC) and open-ended (OE) summed scores on a Wisconsin reading tryout form. θ is standardized; EAP estimates of θ associated with the MC and OE summed scores are shown in the margins. The unshaded area in the table represents the central 99% HDR for the response patterns

Open-Ended (Summed) Rated Score

MC Sum Score	Score	0	1	2	3	4	5	6	7	8	9	10	11	12
	EAPs ↓→	-2.8	-2.5	-2.2	-1.9	-1.6	-1.3	-1.0	-0.6	-0.2	0.1	0.5	1.0	1.4
0	-2.3	-3.2	-3.0	-2.8	-2.6	-2.5	-2.4	-2.2	-2.1	-2.0	-1.9	-1.8	-1.8	-1.7
1	-2.2	-3.1	-2.9	-2.8	-2.6	-2.4	-2.3	-2.1	-2.0	-1.9	-1.8	-1.7	-1.7	-1.6
2	-2.1	-3.0	-2.9	-2.7	-2.5	-2.3	-2.2	-2.0	-1.9	-1.8	-1.7	-1.6	-1.6	-1.5
3	-2.0	-3.0	-2.8	-2.6	-2.4	-2.2	-2.1	-1.9	-1.8	-1.7	-1.6	-1.5	-1.4	-1.4
4	-1.8	-2.9	-2.7	-2.5	-2.3	-2.1	-1.9	-1.8	-1.7	-1.6	-1.5	-1.4	-1.3	-1.2
5	-1.6	-2.8	-2.6	-2.3	-2.1	-1.9	-1.8	-1.6	-1.5	-1.4	-1.3	-1.2	-1.2	-1.1
6	-1.4	-2.7	-2.4	-2.2	-2.0	-1.8	-1.6	-1.5	-1.4	-1.3	-1.2	-1.1	-1.0	-1.0
7	-1.2	-2.5	-2.2	-2.0	-1.8	-1.6	-1.4	-1.3	-1.2	-1.1	-1.0	-1.0	-0.9	-0.9
8	-1.0	-2.2	-1.9	-1.7	-1.5	-1.4	-1.3	-1.2	-1.1	-1.0	-0.9	-0.8	-0.8	-0.7
9	-0.9	-1.9	-1.6	-1.4	-1.3	-1.2	-1.1	-1.0	-0.9	-0.8	-0.8	-0.7	-0.6	-0.6
10	-0.7	-1.5	-1.3	-1.2	-1.1	-1.0	-0.9	-0.8	-0.8	-0.7	-0.6	-0.5	-0.5	-0.4
11	-0.5	-1.2	-1.0	-0.9	-0.8	-0.8	-0.7	-0.6	-0.6	-0.5	-0.4	-0.4	-0.3	-0.2
12	-0.3	-0.8	-0.8	-0.7	-0.6	-0.6	-0.5	-0.5	-0.4	-0.3	-0.2	-0.2	-0.1	-0.0
13	-0.1	-0.6	-0.5	-0.5	-0.4	-0.4	-0.3	-0.3	-0.2	-0.1	-0.0	0.0	0.2	0.3
14	0.5	-0.3	-0.3	-0.3	-0.2	-0.2	-0.1	-0.1	0.0	0.1	0.2	0.3	0.5	0.6
15	0.5	-0.1	-0.1	-0.0	0.0	0.0	0.1	0.2	0.3	0.4	0.5	0.7	0.9	1.1
16	1.1	0.2	0.2	0.2	0.3	0.3	0.4	0.5	0.6	0.7	0.9	1.1	1.4	1.7

(roughly, near the main diagonal of the table), the summed score on each part has a larger "effect" on the scaled score than it may when the scores "disagree." In the latter case, when the scores are inconsistent, the OE score is effectively given less weight because the OE section is less reliable (or discriminating).

Figure 4.1 illustrates the system graphically. The contents of the cell in table 4.1 for $x_{MC}=15$ and $x'_{OE}=9$ are shown expanded at the lower right of the figure: the population OE distribution [$\phi(\theta)$], the likelihood for OE score 9[$L^{OE}_{9}(\theta)$], the likelihood for MC score 15[$L^{MC}_{15}(\theta)$], the product of those three densities, and the likelihood for θ given OE score 9 and MC score 15[$L_{15\&9}(\theta)$]. (To place all the likelihoods on approximately the same scale for the graphic, the population distribution [$\phi(\theta)$], the likelihood for the OE score [$L^{OE}_{x'}(\theta)$], and the likelihood for MC score [$L^{MC}_{x}(\theta)$] have all been normalized to have a maximum ordinate of 1.0 in figures 4.1–4.4)

The plot of the likelihoods in the lower right-hand corner of figure 4.1 is expanded in figure 4.2. Referring to table 4.1, we find that the value of EAP[θ|*x* , *x'*] for the combination is 0.5, while the MC EAP[θ|*x*] (in the row margin of table 4.1) is also 0.5, and the OE EAP[θ|*x'*] (in the column margin of table 4.1) is 0.1. Thus, although the likelihood for the combination appears to be between those for the MC score and the OE score, the "average" computed, using the combination likelihood, is

Table 4.2. The values of $s.d.[\theta|x, x']$ for combinations of MC and OE summed scores on a Wisconsin reading tryout form. θ is standardized; EAP estimates of θ associated with the MC and OE summed scores are shown in the margins. The unshaded area in the table represents the central 99% HDR for the response patterns

Open-Ended (Summed) Rated Score

MC Sum Score	Score	0	1	2	3	4	5	6	7	8	9	10	11	12
	EAPs ↓→	−2.8	−2.5	−2.2	−1.9	−1.6	−1.3	−1.0	−0.6	−0.2	0.1	0.5	1.0	1.4
0	−2.3	0.51	0.52	0.51	0.49	0.47	0.44	0.42	0.40	0.38	0.37	0.36	0.35	0.35
1	−2.2	0.53	0.54	0.53	0.50	0.48	0.45	0.42	0.40	0.39	0.37	0.36	0.35	0.35
2	−2.1	0.55	0.56	0.54	0.52	0.49	0.46	0.43	0.41	0.39	0.37	0.36	0.35	0.34
3	−2.0	0.58	0.58	0.56	0.53	0.50	0.46	0.43	0.41	0.39	0.37	0.36	0.35	0.34
4	−1.8	0.61	0.61	0.59	0.55	0.51	0.47	0.43	0.41	0.38	0.37	0.35	0.34	0.33
5	−1.6	0.65	0.64	0.61	0.56	0.51	0.47	0.43	0.40	0.38	0.36	0.34	0.33	0.32
6	−1.4	0.70	0.68	0.63	0.57	0.51	0.46	0.42	0.39	0.36	0.35	0.33	0.32	0.31
7	−1.2	0.75	0.70	0.63	0.56	0.50	0.44	0.40	0.37	0.35	0.33	0.32	0.31	0.30
8	−1.0	0.80	0.71	0.62	0.54	0.47	0.42	0.38	0.36	0.34	0.32	0.31	0.30	0.30
9	−0.9	0.81	0.68	0.57	0.49	0.44	0.39	0.36	0.34	0.32	0.31	0.31	0.30	0.30
10	−0.7	0.75	0.61	0.51	0.44	0.40	0.37	0.34	0.33	0.32	0.31	0.31	0.30	0.31
11	−0.5	0.63	0.51	0.44	0.40	0.37	0.35	0.33	0.32	0.31	0.31	0.31	0.32	0.33
12	−0.3	0.49	0.43	0.39	0.36	0.35	0.34	0.33	0.32	0.32	0.32	0.33	0.34	0.36
13	−0.1	0.40	0.37	0.36	0.35	0.34	0.34	0.33	0.33	0.34	0.35	0.37	0.39	0.42
14	0.5	0.36	0.36	0.35	0.35	0.35	0.35	0.35	0.36	0.37	0.39	0.42	0.45	0.49
15	0.5	0.37	0.37	0.37	0.37	0.37	0.38	0.39	0.41	0.43	0.45	0.49	0.53	0.59
16	1.1	0.40	0.40	0.40	0.41	0.42	0.43	0.45	0.47	0.50	0.53	0.57	0.62	0.67

Figure 4.1. Graphical illustration of the IRT pattern-of-summed-scores combination system using data from the Wisconsin third-grade reading test. The table in the upper left is a schematic representation of table 4.1. For the cell in table 4.1 for x_{MC}=15 and x'_{OE}=9, the expanded cell at the lower right shows the population distribution $[\phi(\theta)]$, the likelihood for OE score $9[L_9^{OE}(\theta)]$, the likelihood for MC score 15 $[L_{15}^{MC}(\theta)]$, and the product of those three densities, and the likelihood for θ given OE score 9 and MC score $15[L_{15\&9}(\theta)]$

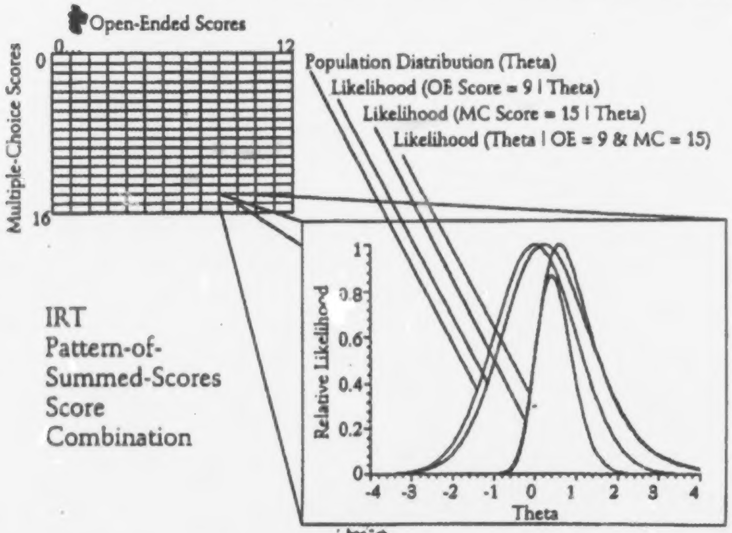
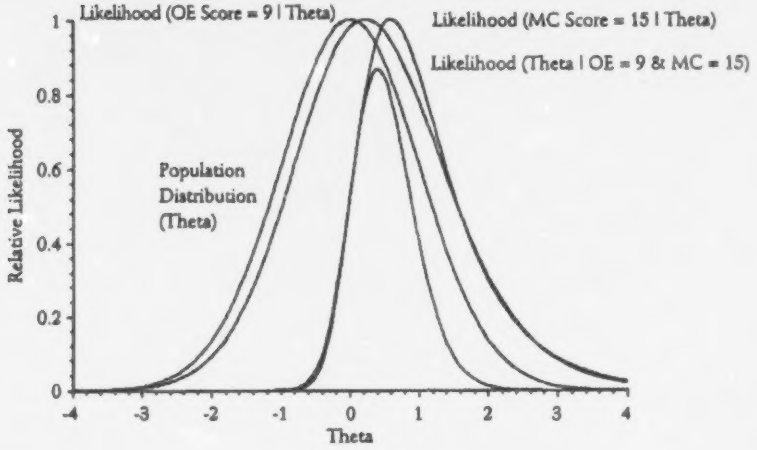


Figure 4.2. The population distribution [$\phi(\theta)$], the likelihood for OE score 9[$L_9^{OE}(\theta)$], the likelihood for MC score 15[$L_{15}^{MC}(\theta)$], the product of those three densities, and the likelihood for θ given OE score 9 and MC score 15[$l_{15\&9}(\theta)$], using data from the Wisconsin third-grade reading test



approximately in the same location as the MC EAP scale score. Table 4.2 gives the value of $s.d.[\theta|x, x'] = 0.45$ for this cell, and the likelihood plotted in figure 4.2 shows this to be a rather accurate description: The inflection points are a little higher than 0 and a little lower than 1.

Figure 4.3 shows a similarly constructed graphic for the combination with MC score 16 and OE score 12—the maximum summed score for each component. In this case, the combination likelihood (the product of both component likelihoods and the population distribution) has ($EAP[\theta|x, x']$) = 1.7 (from table 4.1), while the EAPs for the two components are 1.1 (for the MC section) and 1.4 (for the OE section)—both from the margins of table 4.1. Again, this kind of likelihood-based, score-combination system is better taken as a system for combining evidence (about θ) than as an averaging system, with or without weights. For example, when the examinee obtains the maximum score on both of the two components, the evidence is compounded that the examinee's proficiency is very high—far from the mean of the population distribution.

Figure 4.4 shows the likelihoods for an extremely unlikely combination: all items correct on the MC section (a summed score of 16) and no points on the OE section (summed score 0). This situation presents difficulty for any score combination system. The value of ($EAP[\theta|x]$) for the MC section alone is 1.1, indicating relatively high proficiency, while that for ($EAP[\theta|x']$) for the OE section alone is –2.8, indicating very low proficiency. The value of ($EAP[\theta|x,x']$) for the combination likelihood is 0.2, which is a kind of compromise, but this is a compromise that comes with a warning.

The Probabilities of Score Combinations. The IRT model also gives the probability for each combination of scores x and x'—equation (2). For the score combinations in table 4.1, the values of $P_{xx'}$ range from approximately 0.07 to less than 0.000005; a truncated representation of

Figure 4.3. The population distribution [f(q)], the likelihood for OE score 12[LOE12 (q)], the likelihood for MC score 16[LMC16 (q)], the product of those three densities, and the likelihood for q given OE score 12 and MC score 16[L16&12(q)], using data from the Wisconsin third-grade reading test

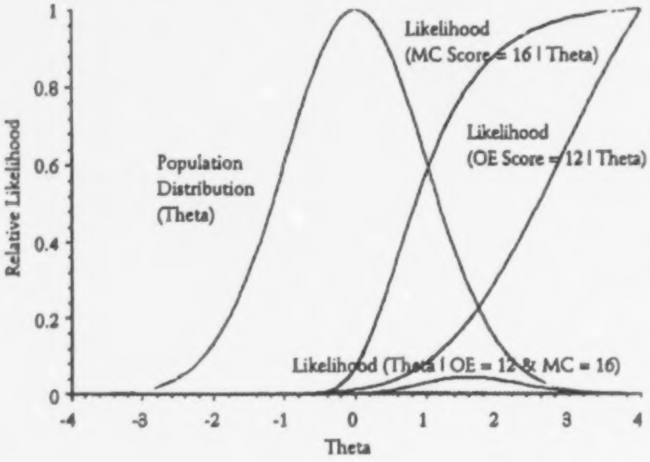
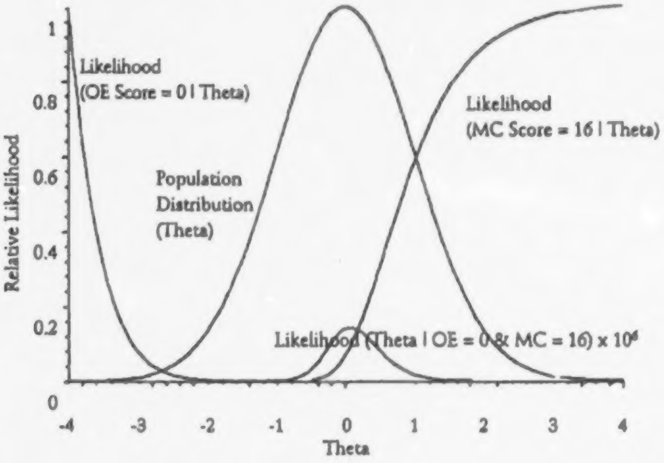


Figure 4.4. The population distribution [$\phi(\theta)$], the likelihood for OE score 0[$L_0^{OE}(\theta)$], the likelihood for MC score 16 [$L_{16}^{MC}(\theta)$], the product of those three densities, and the likelihood for θ given OE score 0 and MC score 16[$L_{16\&0}(\theta)$], using data from the Wisconsin third-grade reading test



those values of $P_{xx'}$ is shown in table 4.3. Considered individually, the values of $P_{xx'}$ are not readily interpretable as reflecting likely or unlikely events in any absolute sense because the magnitude of the individual $P_{xx'}$ depends on the number of row and column score-points. However, the values of $P_{xx'}$ may be used to construct a (1–α)100% "highest density region" (HDR; Novick & Jackson, 1974) for the response combinations.

To construct the HDR, first sort the cells in order of $P_{xx'}$, from largest to smallest, and construct the cumulative distribution of $P_{xx'}$ using that sorted list. As an example, locate the 99% HDR by including the region of all those cells that contribute to the first 99% of the cumulative total of $P_{xx'}$. This region has the properties that include 99% of the modeled response probability (by construction). The probability of any response combination within the region is higher than the probability of any response combination excluded from the region. According to the model, 99% of the examinees should obtain score combinations in that list of cells.

To illustrate, the 99% HDR was located for the Wisconsin reading data, as well as the 99.9% HDR; they are shown with shading in table 4.4. (The same cells are shaded in tables 4.1–4.3). In table 4.4 (as well as tables 4.1–4.3), the 99% HDR is shown with no shading, and the region excluded from the 99.9% HDR is shaded darkly. The light gray shading and the unshaded area together represent the 99.9% HDR. Any response combination in the darkly shaded area in tables 4.1–4.4 is unusual, in that, according to the model, fewer than 1 in 1,000 examinees should produce responses in that region.

Returning our attention to the unlikely score combination illustrated in figure 4.4, we now note that the likelihood for the combination has been multiplied by 10^6 to make it visible on the graphic. The shadings in tables 4.1–4.4 indicate that, according to the IRT model, this particular score combination is one that the model says should occur rarely. Rather than accept any score for this combination (Which one should we accept? The high MC score? The low OE score? An average that represents neither?), this information could be used in some testing systems to indicate that either the test or the model is somehow inappropriate for examinees with this response pattern and that further testing might be useful.

An Aside: Use of $P_{xx'}$ as the Basis of an "Appropriateness Index." For any test that is constructed of blocks of items [MC-OE combinations are just one example; a computerized adaptive test (CAT) that administers testlets sequentially is another], tables of the same general form as table 4.3 may also be constructed of

$$P_{xx'} = \int L_{xx'}(\theta)\, d\theta,$$

for each pair of blocks or its generalization for higher-way tables. These values may be used to construct a (1 – α) 100% HDR for scores for the combination of blocks. These represent the model's predictions of score combinations that are likely and unlikely. Any score combination that lies outside the (1 – α) 100% HDR could be flagged in much the same way that "appropriateness indices" such as l_z (Drasgow, Levine, & Williams, 1985) are used to flag response patterns that are relatively unlikely, according to an IRT model. However, unlike l_z, which relies on a

Table 4.3. The values of $P_{xx'}$ for combinations of MC and OE summed scores on a Wisconsin reading tryout form. Entries in the table are 10,000$P_{xx'}$ truncated, with a leading 0.0 suppressed (that is, 001 is 0.0001). The unshaded area in the table represents the central 99% HDR for the response patterns

Open-Ended (Summed) Rated Score

MC Sum Score	Score	0	1	2	3	4	5	6	7	8	9	10	11	12
	EAPs ↓→	-2.8	2.5	2.2	1.9	-1.6	-1.3	-1.0	-0.6	-0.2	0.1	0.5	1.0	1.4
0	-2.3	000	000	000	000	000	000	000	000	000	000	000	000	000
1	-2.2	000	000	000	000	001	002	002	001	000	000	000	000	000
2	-2.1	000	000	000	002	005	008	009	007	003	001	000	000	000
3	-2.0	000	000	001	004	011	018	021	017	010	004	000	000	000
4	-1.8	000	000	001	006	016	030	038	034	022	009	002	000	000
5	-1.6	000	000	001	007	019	039	054	054	038	018	004	000	000
6	-1.4	000	000	001	006	019	043	067	074	058	030	009	001	000
7	-1.2	000	000	001	005	017	042	074	092	080	046	015	002	000
8	-1.0	000	000	000	003	014	038	075	104	102	064	023	004	000
9	-0.9	000	000	000	002	010	033	072	112	122	086	034	006	000
10	-0.7	000	000	000	001	008	027	066	116	141	112	049	011	000
11	-0.5	000	000	000	001	005	022	060	117	161	144	072	018	001
12	-0.3	000	000	000	000	004	018	054	118	184	188	108	031	003
13	-0.1	000	000	000	000	003	014	048	119	212	253	170	058	007
14	0.5	000	000	000	000	002	011	042	118	245	348	284	120	020
15	0.5	000	000	000	000	001	008	033	108	266	463	482	270	064
16	1.1	000	000	000	000	000	004	019	072	218	482	675	542	196

Gaussian approximation for the distribution of the log-likelihood for its questionable *p*-values (Reise & Flannery, 1996), the probability statements associated with use of a (1 – α) 100% HDR for two- (or three- or four-) way classifications of the item responses blockwise may be computed directly from the IRT model.

What should be done with examinees whose responses are flagged as unlikely according to the model? This question raises difficult policy questions, and its answer certainly depends on the purpose of the test. The mismatch between the examinee's performance on one block and another may mean the examinee cheated on one block, in which case the better measure may be their lower performance, or it may be that something else went wrong with their performance on the block on which they scored lower (distraction? computer difficulties?), in which case the higher of the two scores may be more valid. The context of high-stakes testing, which is expensive in both time and money for the examinee, might add further considerations. Davis and Lewis (1996) suggest several possible courses of action that could be followed if the test was computerized: One set of possible actions includes an on-line extension of the test, either switching from a CAT system to a long linear form or using a special block of "silver bullet items" to estimate more accurately the

Table 4.4. The 99% and 99.9% HDRs for score combinations on a Wisconsin reading tryout form. EAPs associated with the MC and OE summed scores are shown in the margins

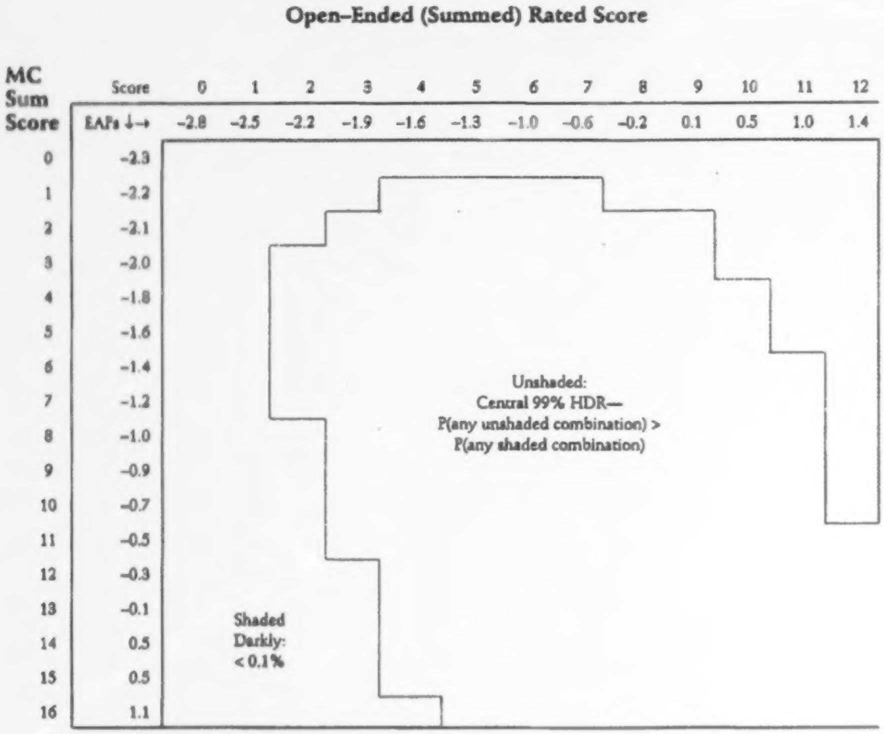

Open-Ended (Summed) Rated Score
MC Sum Score
Score
0 1 2 3 4 5 6 7 8 9 10 11 12
EAPs ↓→
-2.8 -2.5 -2.2 -1.9 -1.6 -1.3 -1.0 -0.6 -0.2 0.1 0.5 1.0 1.4
0 -2.3
1 -2.2
2 -2.1
3 -2.0
4 -1.8
5 -1.6
6 -1.4
7 -1.2
8 -1.0
9 -0.9
10 -0.7
11 -0.5
12 -0.3
13 -0.1
14 0.5
15 0.5
16 1.1
Unshaded:
Central 99% HDR—
P(any unshaded combination) >
P(any shaded combination)
Shaded
Darkly:
< 0.1%

Figure 4.5. The values of $EAP[\theta|x,x']$, computed using the 3PL/GR model combination, plotted as a surface of points over a grid representing the OE and MC summed scores, using data from the Wisconsin third-grade reading test (Only the points for the central 95% HDR of the score combinations are plotted)

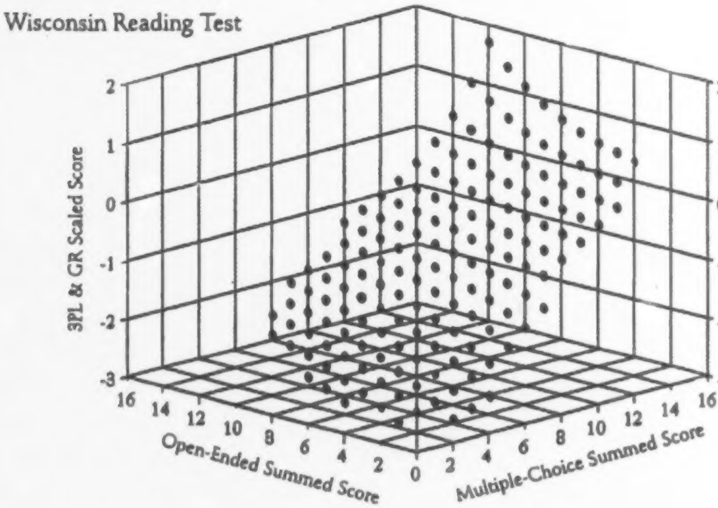
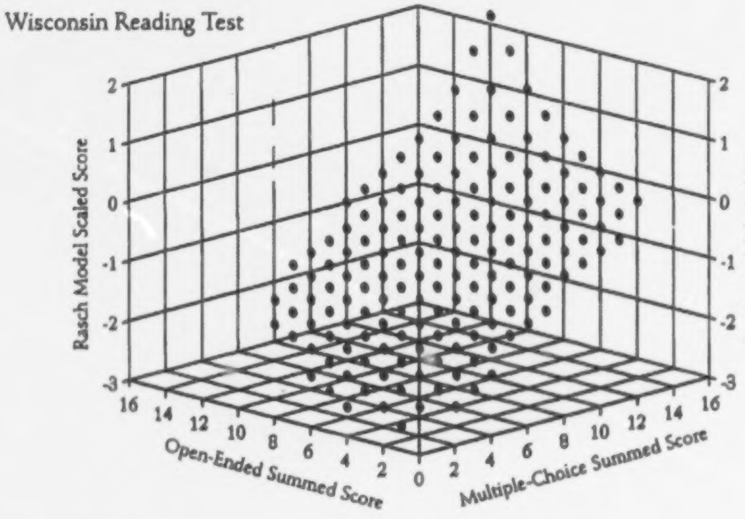


Figure 4.6. The values of $\text{MAP}[\theta|x,x']$, computed using the Rasch model, plotted as a surface of points over a grid representing the OE and MC summed scores, using data from the Wisconsin third-grade reading test (Only the points for the central 95% HDR of the score combinations are plotted)

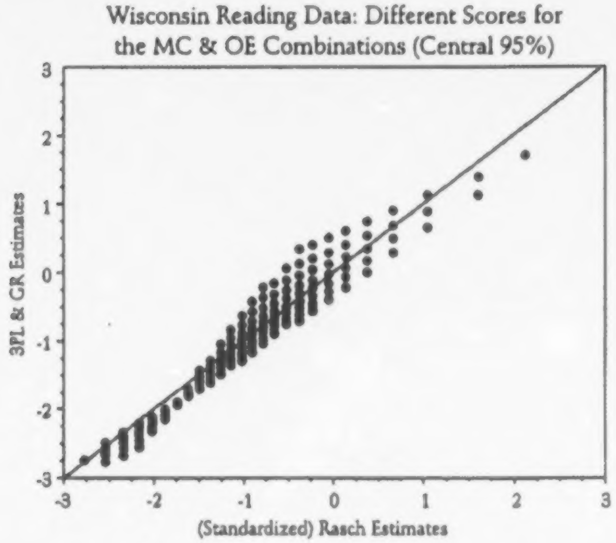


examinee's proficiency. Other possible actions include score cancellation and retesting.

The Relation of Likelihood-Based Score Combination with Weighted Linear Combinations and with Rasch Model Scores. Figure 4.5 shows the score combination values of $\text{EAP}[\theta|x, x']$ for the Wisconsin third-grade reading test plotted as a point-surface over a grid representing the MC and OE summed scores. (Only the points for the central 95% HDR of the score combinations are plotted.) The curvature of that surface represents the difference between this score-combination system and a linear-weighted average of the two scores. The corresponding points for any linear-weighted average system would lie on a plane. IRT "adjustments" for the relative difficulty and discrimination of the MC and OE items, done on a point-by-point basis as the likelihoods are used to combine the evidence about proficiency each represents, cannot be exactly matched by any linear-weighting scheme. However, the fact that the rows of points rise more quickly in the direction of increase of the MC score than they do in the direction of increase for the OE score means that the IRT system is effectively "weighting" the MC "points" more than the OE "points."

Analysis of data like these with a Rasch-family model produces scale scores that are the same for any score combination that yields the same total summed score (Masters & Wright, 1984). In the case of Rasch-family models, the two-way array of values of $\text{EAP}[\theta|x, x']$, such as is shown in table 4.1, is superfluous; any combination of x and x' that have the same total score have the same scale score. When applied to data that the 3PL/GR model combination fits with different discrimination values for the MC items (as a set) and the OE items

Figure 4.7. The values of EAP[$\theta|x,x'$], computed using the 3PL/GR model combination, plotted against the values of MAP[$\theta|x,x'$], computed using the Rasch model, using data from the Wisconsin third-grade reading test. (Only the points for the central 95% HDR of the score combinations are plotted)



(as a set), Rasch-family model scale scores and the 3PL/GR model combination scale scores differ somewhat. Figure 4.6 shows the values of Rasch-family MAP[$\theta|x, x'$] as computed with the computer program Bigsteps (Wright & Linacre, 1992) for the Wisconsin third-grade reading test plotted as a point-surface over a grid representing the MC and OE summed scores. (Again, only the points for the same central 95% HDR of the score combinations are plotted as in figure 4.5; the computation of the HDR is based on the 3PL/GR model.) Unlike the pattern shown in figure 4.5, the Rasch-family scale scores increase at exactly the same rate as the MC scores increase as they do for an OE score increase.

Figure 4.7 shows the 3PL/GR score-combination values of EAP[$\theta|x, x'$] plotted against the Rasch-family MAP[$\theta|x, x'$] estimates. (Because the Rasch-family estimates are originally computed on a different scale, their values in figure 4.7 have been standardized to have the same mean and variance as the values of EAP[$\theta|x, x'$]). We see in figure 4.7 that some score combinations for which the Rasch-family values of MAP[$\theta|x, x'$] and the 3PL/GR values of EAP[$\theta|x, x'$] differ fairly substantially. Because the 3PL/GR scale scores are computed to account for guessing on MC items (which almost certainly happens) as well as different relative values of the "points" for the number-correct MC score as opposed to the rated OE score (which also almost certainly contains some truth), we tend to believe that when the two scores differ, the 3PL/GR scores may be more valid.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Davis, L. A., & Lewis, C. (1996). *Person-fit indices and their role in the CAT environment.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, April 9–11.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67–86.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529–544.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill Book Company.

Reise, S. P., & Flannery, W. P. (1996). Assessing person-fit on measures of typical performance. *Applied Psychological Measurement, 9,* 9–26.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, 17.*

Thissen, D. (1991). *Multilog user's guide—Version 6* [Computer program]. Chicago: Scientific Software, Inc.

Wright, B. D., & Linacre, J. M. (1992). *Bigsteps Rasch analysis* [Computer program] Chicago: MESA Press.

SECTION 5

Enhancing the Validity of NAEP Achievement Level Score Reporting

Ronald K. Hambleton University of Massachusetts, Amherst

August 1997

Enhancing the Validity of NAEP Achievement Level Score Reporting*

The National Assessment of Educational Progress (NAEP) provides policymakers, educators, and the public with information about the reading, mathematics, science, geography, history, and writing knowledge and skills of elementary, middle, and high school students. NAEP also monitors changes in student achievement over time. NAEP uses considerable statistical and psychometric sophistication in its test design, data collection, test data analysis, and scaling (Beaton & Johnson, 1992; Johnson, 1992; Mislevy, Johnson, & Muraki, 1992). In fact, NAEP may be the most technically sophisticated assessment system in the world.

Less attention, however, appears to be given to the ways in which the complex NAEP data are organized and reported. In fact, the contrast between the efforts and success in producing sound technical NAEP instruments, drawing samples, administering the assessments, and analyzing the assessment data and the efforts and success in disseminating NAEP results is striking.

Concerns about NAEP data reporting have become an issue in recent years. These concerns have been documented by Hambleton and Slater (1995, in press), Jaeger (1992), Koretz and Deibert (1993), Linn and Dunbar (1992), and Wainer (1996, 1997a, 1997b). The objectives of this paper are as follows:

(a) provide background on NAEP score reporting with achievement levels (since 1990);

(b) describe the results of a small-scale study of the understandability of NAEP score reports among policymakers; and

(c) review several promising new directions in score reporting along with their implications for NAEP—redesign of NAEP displays (Wainer, Hambleton, & Meara, in progress), guidelines for preparing displays (Hambleton, Slater, & Allalouf, in progress), and market-basket reporting (the idea was suggested by Mislevy, Bock, & Thissen).

In fact, NAEP may be the most technically sophisticated assessment system in the world.

Background on NAEP Score Reporting

"What is the meaning of a NAEP mathematics score of 220?" "Is a national average of 245 in mathematics good or bad?" These two questions were posed by policymakers and educators in a study conducted in 1994 by Hambleton and Slater (1995, in press) following the release of the Executive Summary of the 1992 NAEP national and state mathematics results. Questions about the meaning of scores are also frequently asked by those attempting to make sense of IQ, SAT, ACT, and other standardized achievement test scores. People are more familiar with popular ratio scales, such as those used in measuring distance, time, and weight, than with educational and psychological test score scales.

---

*See also Laboratory of Psychometric and Evaluative Research Report No. 317. Amherst, MA: University of Massachusetts, School of Education.

Test scores are elusive. Even the popular percentage score scale, which many think they understand, cannot be understood unless the domain of content to which percentage scores are referenced is clear and the method used for selecting assessment items is known. Few seem to realize the importance of these two critical pieces of information in interpreting percentage scores. This problem is also present in state legislation being written that establishes a passing score on an important statewide test, without detailed knowledge of the test's content or difficulty.

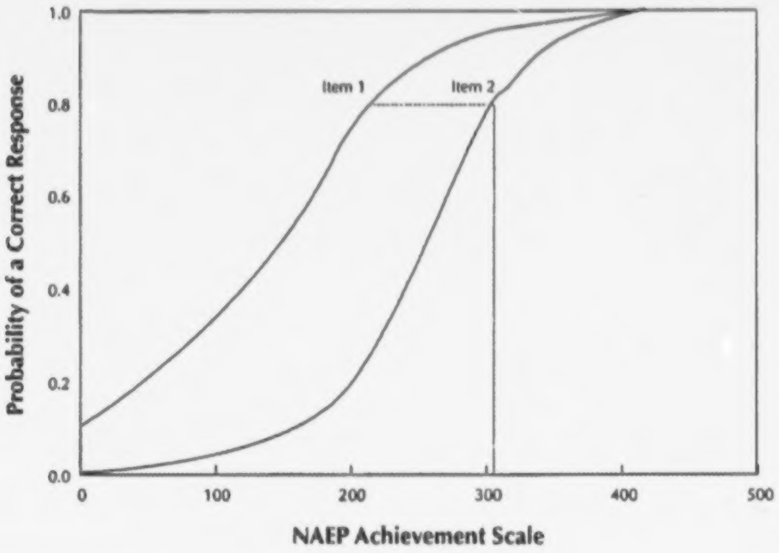
One way to make statistical results more meaningful for intended audiences is to connect them to numbers that may be better understood than test scores and test score scales.

One solution to the score interpretation problem is simply to interpret the scores in a normative way (i.e., scores obtain meaning or interpretability by being referenced to a well-defined norm group). All popular norm-referenced achievement tests use norms to assist in test score interpretations. However, normative statements are not always valued. Sometimes the important question policymakers have concerns level of accomplishment—for example, what percentage of students have reached a level of 250 on the assessment? Many policymakers would like to choose points such as 250 to represent well-defined levels of accomplishment (which might be called Basic, Proficient, and Advanced) and then determine the numbers of students in interest groups (e.g., regions of the country) who achieve these accomplishment levels. This is known as criterion-referenced (CR) assessment. Most national and state assessments are examples of CR assessments, and with these assessments, scores need to be interpreted in relation to content domains, anchor points, and/or performance standards (Hambleton, 1994).

NAEP constructed an arbitrary scale with scores ranging from 0 to 500 for each subject area. The scale was obtained as follows: The distributions of scores from nationally representative samples of 4th-, 8th- and 12th-grade students were combined and scaled to a mean of 250 and a standard deviation of approximately 50 (Beaton & Johnson, 1992). The task was then to facilitate CR score interpretations on this scale (Phillips et al., 1993). Placing benchmarks such as grade-level means, state means, and performance of various subgroups of students (e.g., males, females, Black, Hispanic) is helpful in giving meaning to the scale, but these benchmarks provide only a norm-referenced basis for score interpretations.

One way to make statistical results more meaningful for intended audiences is to connect them to numbers that may be better understood than test scores and test score scales. For example, when many persons were concerned recently about flying after the TWA (Flight 800) crash, the airlines reported that only one plane crash occurs for every 2 million flights. In case the safety of air travel still was not clear, the airlines reported that a person could expect to fly every day for the next 700 years without an accident. Most likely, some people felt more confident after hearing these statistics reported in an understandable fashion. Nevertheless, knowing that the

Figure 5.1. Graphical display of two item characteristic curves, item 1 being easier for the examinees than item 2. This display highlights the potential use of item mapping to enhance the meaning of the NAEP scale



Source: Hambleton & Slater, 1995

probability of being in a plane crash is less than 0.0000005 is not very meaningful to most people.

Concerning the reporting of NAEP results, what, for example, does a single point mean? It was noted that the typical student (one at the 50th percentile) gained approximately 48 points between fourth and eighth grades in mathematics (Hambleton & Slater, 1995), which converts to approximately 1.2 points per month of instruction (a gain of 48 points over 40 months of instruction). Recognizing that the growth over 4 years is not necessarily linear (see, for example, grade-equivalent scores on standardized achievement tests), it could be said that 1 point is at least roughly equivalent to 1 month of regular classroom instruction. Secretary of Education Richard Riley used this approach recently to communicate findings in the 1996 NAEP Science Assessment, and the connection between NAEP score points and instructional time appeared to be a valuable way to relay the meaning of points on the NAEP scale.

Other possibilities with considerable promise for CR interpretations of scores include item mapping, anchor points, performance standards (called "achievement levels" in the NAEP context), and benchmarking (Phillips et al., 1993). These approaches capitalize on the fact that scales based on item response theory (IRT) locate both the assessment material and the examinees on the same reporting scale. Thus, at any particular point (i.e., ability level) of interest, the types of tasks that examinees at that ability level can successfully complete can be determined. Tasks that these examinees cannot complete with some stated degree of accuracy (e.g., 50% probability of successful completion) can also be identified. Descriptions at these points of interest can be developed and exemplary items selected—that is, items to highlight what examinees at these points of interest might be expected to accomplish (Bourque, Champagne, & Crissman, 1997; Mullis, 1991).

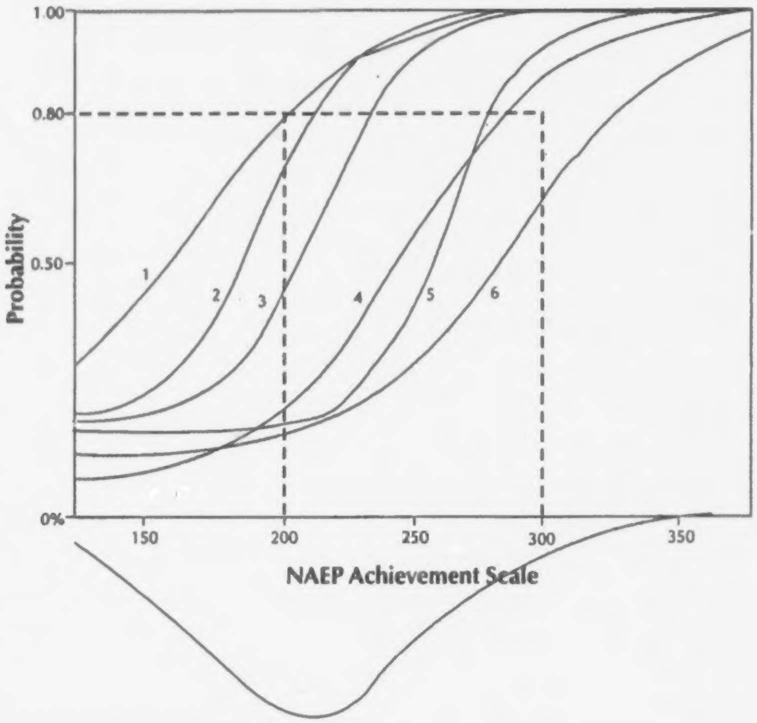
Figure 5.1 shows the "item characteristic curves" for two NAEP dichotomously scored items (Hambleton, Swaminathan, & Rogers, 1991). At any point on the NAEP achievement (i.e., proficiency) scale, the probability of a correct response (i.e., answer) can be determined. Item 2 is the more difficult item since, regardless of ability, the probability of a correct response to item 2 is lower than item 1. The location on the ability scale at which an examinee has an 80% probability of success for an item is called the "RP_{80}" for the item. In figure 5.1, the RP_{80} for item 1 is estimated at 210 and the RP_{80} for item 2 is approximately 306. This is known as "item mapping." Each item in a NAEP assessment can be located on the NAEP achievement scale according to RP_{80} values. If 80% probability is defined as the probability at which an examinee can reasonably be expected to know or accomplish something (other probabilities, such as 65%, have often been used), then an examinee with an ability score of approximately 210 could be expected to answer items such as item 1 and other items with RP_{80} values of approximately 210 on a fairly consistent basis (i.e., approximately 80% of the time). In this way, a limited type of CR interpretation can be made even though examinees with scores of approximately 210 may never have actually been administered item 1 or other items similar to it as part of their assessment.

The validity of CR interpretations depends on the extent to which a unidimensional reporting scale fits the data to which it is applied. If a group of examinees scores 270, a score of 270 is then made more meaningful by describing the contents of items such as those with RP_{80} values of approximately 270. The item-mapping method is one way to facilitate CR interpretations of points on the NAEP scale or any other scale to which items have been referenced. Cautions regarding this approach have been clearly outlined by Forsyth (1991). A major concern is the nature of the inferences that can legitimately be made from predicted examinee performance on a few test items.

A variation on the item-mapping method is to select arbitrary points on a scale and then to describe these points thoroughly through the knowledge and skills measured by items with RP_{80} values in the neighborhood of the selected points. In the case of NAEP reporting, arbitrarily selected points have been 150, 200, 250, 300, and 350. The item-mapping method can then be used to select items that can be answered correctly by examinees at those points. For example, using the item characteristic curves reported in figure 5.2, at 200, items such as 1 and 2 could be selected. At 250, item 3 would be selected. At 300, items 4 and 5 would be selected. At 350, item 6 would be selected. Of course, in practice, many items might be available for making selections to describe the knowledge and skills associated with performance at particular points along the ability scale.

Currently, RP_{65} values, rather than RP_{80} values, are used by NAEP, and items that clearly distinguish between anchor points are preferred when describing those points. For more details on current practices, see Beaton and Allen (1992), Mullis (1991), and Phillips et al. (1993). Note, too, that a similar process can be implemented for the individual score points on polytomously scored tasks. The method is not limited to dichotomously scored response data.

Figure 5.2. Graphical display highlighting the use of anchor points and item characteristic curves to enhance the meaning of the NAEP scale



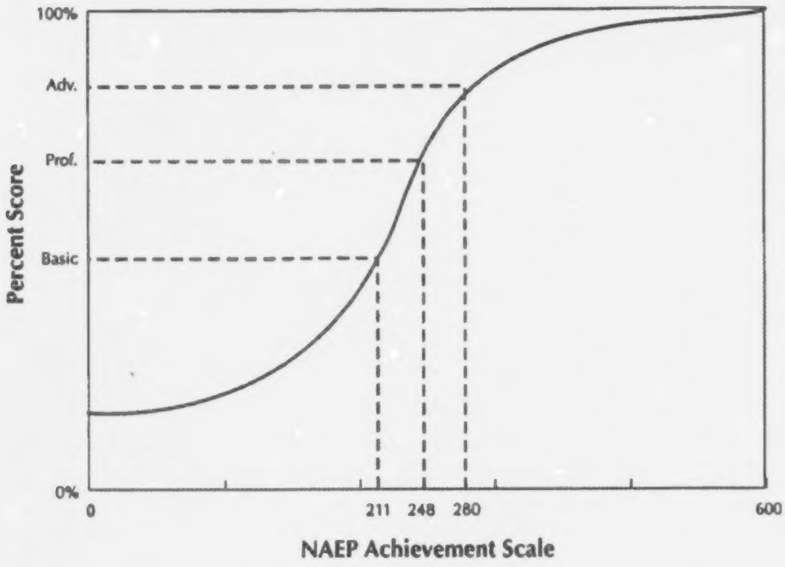
Source: Hambleton & Slater, 1995

The National Assessment Governing Board (NAGB) was not completely satisfied with the use of arbitrary points (i.e., anchor points) for reporting NAEP results. One reason was that the points 200, 250, and 300 became incorrectly associated by the media and policymakers with the standards of performance expected of 4th-, 8th-, and 12th-grade students, respectively. To eliminate the confusion, as well as to respond to the demand from some policymakers and educators for real performance standards on the NAEP scale, NAGB initiated a project to establish performance standards on the 1990 NAEP Mathematics assessment (Hambleton & Bourque, 1991) and has conducted similar projects to set performance standards on NAEP assessments in 1992, 1994, and 1996.

Figure 5.3 depicts the way in which performance standards (set on the test score metric, a scale more familiar to standard-setting panelists than the NAEP achievement scale) can be mapped or placed on the NAEP achievement scale using the test characteristic curve (TCC). (In general terms, the TCC is a weighted average item characteristic curve for items that make up the assessment.) Of course, almost nothing is simple with NAEP, so figure 5.3 is an oversimplification of how the mapping is actually done. However, figure 5.3 depicts how mapping is performed in many state assessments.

The performance standards for a particular grade on the NAEP achievement scale can be used to report and interpret the actual performance of the national sample or any subgroup of interest. With these standards in place, the percentage of students in each performance

Figure 5.3. Graphical display highlighting the connection between performance standards on the test score scale and the NAEP score scale



Source: Hambleton & Slater, 1995

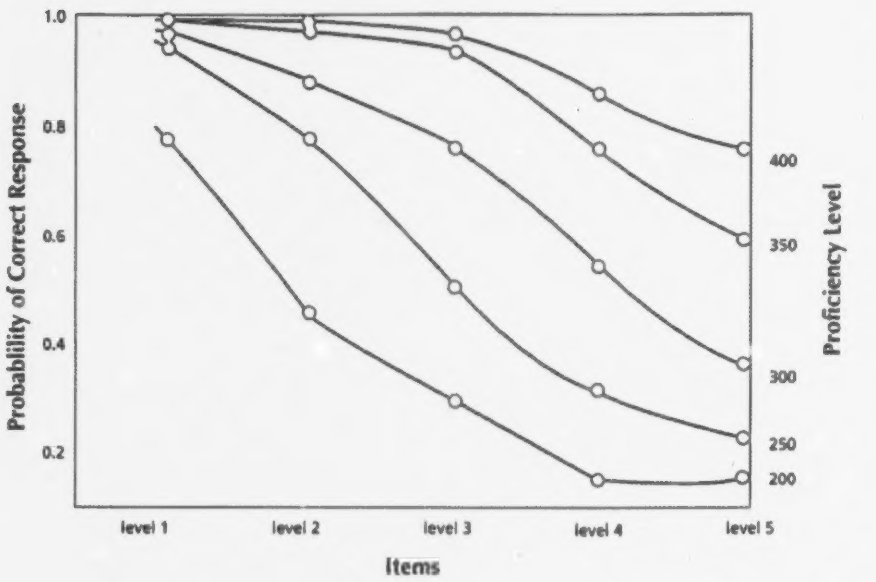
category in score distributions of interest can be reported, and the changes in these percentages can be monitored over time.

Anchor points and performance standards are placed on an achievement scale to enhance the content meaning of scores and to facilitate meaningful CR interpretations of the results. (For example, What percentage of fourth-grade students in the 1996 NAEP Science Assessment are able to perform at the Proficient level or above?) In recent years, both anchor points (e.g., 150, 200, 250, 300, and 350) and performance standards (e.g., borderline scores for Basic, Proficient, and Advanced students in grades 4, 8, and 12) have been placed on NAEP scales. Many states have adopted similar techniques for score reporting.

Performance standards are more problematic than anchor points because they require a fairly elaborate process to establish (e.g., the current design calls for 30 panelists at a grade level working for 5 days) and validate. At the same time, performance standards appear to be greatly valued by many policymakers and educators. For example, many state departments of education use performance standards in reporting, and many states involved in the NAEP trial state assessment have indicated a preference for standards-based reporting over anchor points-based reporting.

Figure 5.4 provides a final example of score reporting. (For additional references, see the report by Phillips et al., 1993.) This example is taken from the National Adult Literacy Survey (Kirsch et al., 1993). Each monotonically decreasing curve represents the performance of adults located at proficiency levels 400, 350, 300, 250, and 200, respectively, on assessment material organized into levels of difficulty (level 1 to level 5). With information about the types of items placed at each level, differences in performance among adults at

Figure 5.4. Average probabilities of correct responses to items along the document scale by adults with different proficiency levels



Source: Kirsch, Jungeblut, Jenkins, & Kolstad, 1993

ability levels 200, 250, 300, 350, and 400 can be better understood. If, instead of choosing adults at particular points (anchor points methodology), adults could be sorted into Below Basic, Basic, Proficient, and Advanced performance categories, then figure 5.4 could be used in standards-based reporting to provide a better understanding of the nature of the performance differences among these four groups of adults.

Small-Scale Study of the Understandability of NAEP Score Reports

The design of tables, figures, and charts to transmit statistical data to enhance their meaningfulness and understandability is a new area of concern in education and psychology (Wainer, 1992; Wainer & Thissen, 1981). However, an extensive literature exists that appears relevant to the topic of data reporting in the fields of statistics and graphic design (Cleveland, 1985; Henry, 1995; Wainer, 1997c).

How bad, or good, is the current situation? Do policymakers and educators understand what they are reading about student achievements and changes over time? Do they make reasonable inferences and avoid inappropriate ones? What is their opinion about the information they are given? Is it important to them? What do they understand and what deficiencies and strengths exist relative to NAEP reports? In view of the shortage of available evidence about the extent to which intended NAEP audiences understand and can use NAEP reports, a small research study was performed by Hambleton and Slater (1995, in press) to investigate the extent to which NAEP Executive Summary Reports were understood by policymakers and educators, to determine the extent to which problems were identified, and to offer a set of recommendations for improving reporting practices. The Technical Review Panel initiated the project, which was funded by the National Center for Education Statistics (NCES).

The 59 participants in the interviews made up a broad audience, similar to the intended audience of NAEP Executive Summary Reports. Persons at state departments of education, attorneys, directors of companies, state politicians and legislative assistants, school superintendents, education reporters, and directors of public relations were among those interviewed.

One problem in understanding the text was due to the use of statistical jargon ...

The interviews were based on the *Executive Summary of the NAEP 1992 Mathematics Report Card for the Nation and the States* (Mullis, Dossey, Owen, & Phillips, 1993). This particular report was chosen because it was relatively brief and was intended to stand on its own merits for policymakers and educators. NAEP Executive Summary Reports are also well known and widely distributed to people working in education or interested in education. (More than 100,000 copies of each Executive Summary are produced.) Furthermore, NAEP Executive Summary Reports, which include both national and state results, are thought to be of interest to the interviewees, who were from different areas of the country.

The goal of the interviews was to determine how much of the information in the Executive Summary Reports was understood. An attempt was made to pinpoint the aspects of reporting readers found confusing and to identify changes that interviewees found would improve their understanding of the results.

The 1992 NAEP Mathematics Executive Summary Report consists of six sections that highlight findings from different aspects of the assessment. Interview questions were designed for each section to ascertain the kind of information interviewees were obtaining from the report. Interviewees were asked to read a brief section of the report and then were questioned on the general meaning of the text or on the specific meaning of certain phrases. Interviewees also examined tables and charts and were asked to interpret some of the numbers and symbols. Interviewees were encouraged to volunteer their opinions and suggestions.

The sample of interviewees was mainly white and included more females (64%) than males (36%). Interviewees were from various areas of the education field, and two education reporters took part in the study. All interviewees indicated that they had medium to high interest in national student achievement results. Most interviewees (90%) were familiar with NAEP in at least a general way, and 64% had read NAEP publications prior to the interview. Almost half the sample had taken more than one course in testing or statistics (46%); one fourth had taken only one course, and another one fourth had taken none.

Nearly all the interviewees (92%) demonstrated a general understanding of the main points of the text summarizing the major findings of the report, though several interviewees commented that they would have liked more descriptive information (e.g., concrete examples). One problem in understanding the text was due to the use of statistical jargon (e.g., statistical significance, variance). This language choice confused and intimidated a number of the interviewees. Several interviewees suggested that a glossary of basic terms would be very helpful. Terms such as Basic, Proficient, Advanced, standard errors, and the NAEP scale could be included in such a glossary.

One example indicates that the phrase "statistically significant" was unclear to many interviewees (42%). Interviewees were expected to know that "statistically significant increases" are not increases resulting from chance events. Fifty-eight percent said that they thought they knew the meaning, but many of the interviewees in this group could not explain what the term meant or why it was used. This was surprising since more than half the interviewees had taken statistics courses. Typical responses to the question "What does statistically significant mean?" were:

- "More than a couple of percentage points."
- "10 percentage points."
- "At least a 5-point increase."
- "More than a handful—you have enough numbers."
- "Statisticians decide it is significant due to certain criteria."
- "The results are important."
- "I wish you hadn't asked me that. I used to know."

The common mistake was to assume "statistically significant differences" were "big and important differences."

Several interviewees mentioned that although they realized that certain terms (e.g., standard error, estimation, confidence level) were important to statisticians, these terms were meaningless to them. They said that their eyes tended to glaze over when technical terms were used in reports or they formed their own working definitions such as those offered above.

Table 1 of the Executive Summary Report is one of the most important and contains a wealth of information. Results in the table are reported for the following categories: grades 4, 8, and 12; 1990 and 1992; average proficiency; and each performance category. Standard errors are given for all statistics in the table. However, many problems with this table were identified in the research study. For example, confusion about the meaning of "At or Above" was seen in table 1. When asked what the 18% in line 1 of table 1 meant (18% of grade 4 students in 1992 were in the Proficient or Advanced category in mathematics), more than half (53%) of the interviewees responded incorrectly. Several interviewees did not look at the table closely enough to see the "Percentage of Students At or Above" heading above the levels. The fact that categories were arranged from Advanced to Basic complicated the use of the table and the concept of "At or Above." In this case, "At or Above" meant summing from right to left, which seemed backward to interviewees when the correct interpretation was given to them.

The common mistake was to assume "statistically significant differences" were "big and important differences."

A summary of six problems that arose when interviewees read table 1 follows:

1. Interviewees were confused by the reporting of average proficiency scores. (Few understood the 500-point NAEP scale.) Proficiency as measured by NAEP and reported on its scale was confused with the category of "proficient students."

2. Interviewees were also baffled by the standard error beside each percentage. The reporting of the standard error interfered with reading the percentages, and the footnotes did not clearly explain what a standard error is and how it could be used.
3. The < and > signs were misunderstood or ignored by most interviewees. Even after reading the footnotes, many interviewees indicated that they were still uncertain about the meaning of the signs.
4. The most confusing point was the reporting of students "At or Above" each proficiency category. Interviewees interpreted these cumulative percentages as the percentage of students in *each* proficiency category. They were surprised and confused when the sum of percentages across any row far exceeded 100%. Contributing to the confusion in table 1 was the presentation of the categories in reverse order of what was expected (i.e., Below Basic, Basic, Proficient, and Advanced). This information as presented required reading from right to left instead of the more common left to right. Only approximately 10% of the interviewees were able to make the correct interpretations of the percentages in table 1.

One frequently offered suggestion is the recommendation to make the reports accessible to nonstatisticians.

5. Footnotes were not always read by interviewees and were often misunderstood when they were read.
6. Interviewees expressed confusion because of variations between NAEP reports and their own state reports.

Given the major difficulties they had in understanding the information contained in table 1, it is not surprising that nearly 80% of the interviewees reported that this table "needs work." This was of concern because table 1 is the penultimate table in the Executive Summary of NAEP results. Several interviewees expressed that bar graphs would have improved the document. More than 90% of those interviewed indicated that they did not have a lot of time to interpret these complex tables, and they believed that a simple graph could be understood relatively quickly.

A second example from Hambleton and Slater may also be informative. Table 2 of the NAEP Executive Summary Report was unclear to approximately 30% of the interviewees. "Cutpoint" and "scale score," examples of NAEP jargon, were the source of the confusion. The interviewees had no idea of the meaning of the numbers in the table, and this information was not contained in the report.

The interviewees made fundamental mistakes in interpreting the figures and tables in the Executive Summary. Nearly all were able to understand the text in general terms, though many would have liked more descriptive information (e.g., definitions of measurement and statistical jargon as well as concrete examples). The problems in understanding the text involved the use of statistical jargon. This confused and even intimidated some interviewees. Some mentioned that, although these terms are important to statisticians, such terms are meaningless to them. After years of seeing these terms in reports, they simply passed over the words in their reading.

Many interviewees offered helpful and insightful opinions about the report. One frequently offered suggestion is the recommendation to make the reports accessible to nonstatisticians. Another comment made by several interviewees was that the report appeared to be "written by statisticians, for statisticians." To remedy this, many suggested removing the statistical jargon. Phrases such as "statistically significant" did not hold much meaning for the policymakers and educators interviewed.

The conclusions and recommendations from a study such as this one must be limited because of its modest nature (only 59 interviews were conducted), the nonrepresentativeness of the persons interviewed (though it was an interesting and important group of policymakers and educators), and the use of only one NAEP report in the study. Note, too, that the research was conducted on a 1992 NAEP report. Reports from 1994 and 1996 appear to be designed better and more responsive to the needs of the intended audiences.

Despite the limitations, several conclusions and recommendations were offered by Hambleton and Slater:

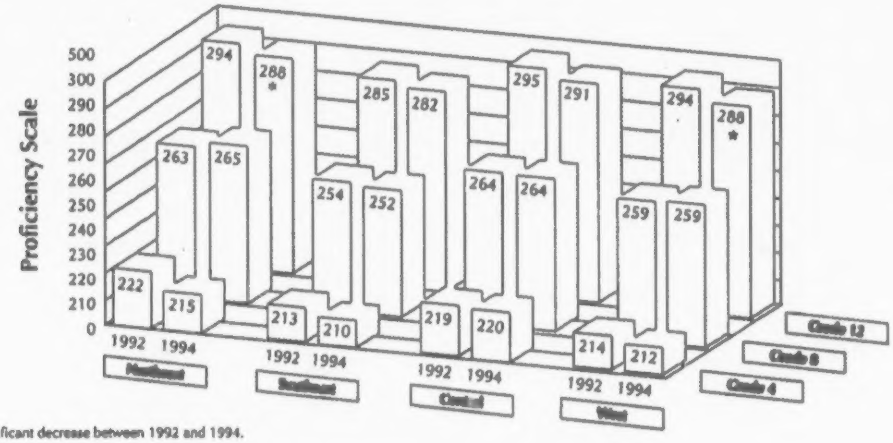
- A considerable amount of misunderstanding was evident concerning the results reported in the 1992 NAEP Mathematics Assessment Executive Summary Report.
- Improvements should include the preparation of a substantially more user-friendly report with simplified figures and tables.
- Reports should be straightforward, short, and clear because of the time constraints experienced by those likely to read these executive summaries.

On the basis of the findings from their study, Hambleton and Slater offered several reporting guidelines for NAEP and state assessments:

1. Make charts, figures, and tables understandable without reference to the text. (Readers did not seem willing to search through the text for interpretations.)
2. Field-test graphs, figures, and tables on focus groups representing the intended audiences. (Many important problems can be identified from field-testing report forms. The situation is analogous to field-testing assessment materials prior to their use.)
3. Ensure that charts, figures, and tables can be reproduced and reduced without loss of quality. (Because interesting and important results will be copied and distributed, copies must be legible.)
4. Keep graphs, figures, and tables relatively simple and straightforward to minimize confusion and shorten the time required by readers to identify the main trends in the data.
5. NAEP Executive Summaries should include an introduction to NAEP and NAEP scales. A glossary should also be provided. Statistical jargon should be deemphasized; tables, charts, and graphs should be simplified; and more boxes and graphics should be used to highlight the main findings.
6. Specially designed reports may be needed for each intended audience. For example, policymakers might find short reports with bulleted text that highlights main points such as conclusions helpful.

Ways must be found to balance statistical rigor with the informational needs, time constraints, and quantitative literacy of intended audiences.

Figure 5.5. Average reading proficiency by grade and by region—NAEP 1992 and 1994



*Significant decrease between 1992 and 1994.

Source: Williams, Reese, Campbell, Mazzeo, & Phillips, 1995

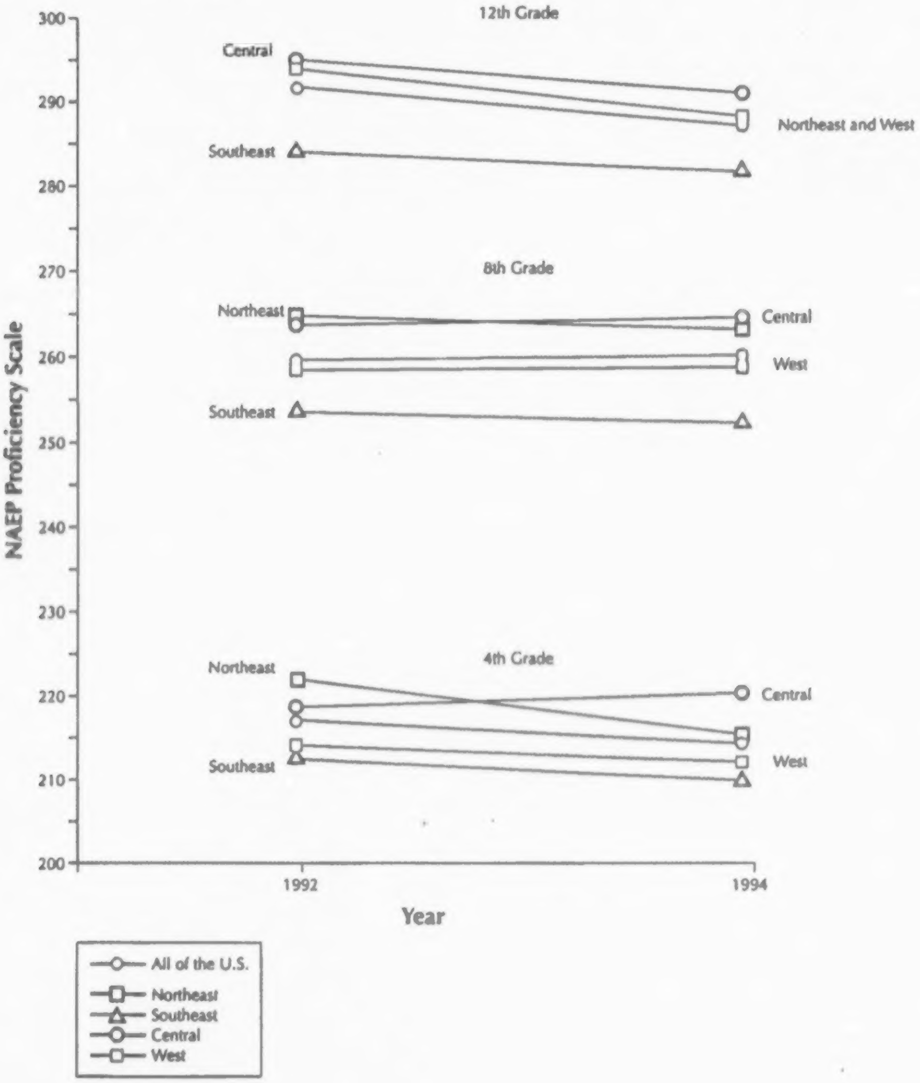
The National Adult Literacy Survey (Kirsch et al., 1993), conducted by NCES, Westat, and the Educational Testing Service, and the recently published standards-based report on the 1996 NAEP Science Assessment (Bourque, Champagne, and Crissman, 1997) appear to have benefited from some of the earlier evaluations of NAEP reporting. They provide excellent examples of data reporting, with figure 5.4 being one such example. A broad program of research, involving measurement specialists, graphic design specialists (Cleveland, 1985), and focus groups representing intended audiences for reports, is needed to build on the successes represented in the reports by Bourque, Champagne, and Crissman (1997), Jaeger (1992), Kirsch et al. (1993), Koretz and Deibert (1993), and Wainer (1996, 1997a, 1997b). Ways must be found to balance statistical rigor with the informational needs, time constraints, and quantitative literacy of intended audiences. Examples from this emerging research are described in the following sections of the paper.

Three Current Advances in Score Reporting

Wainer, Hambleton, and Meara Study

NCES recently funded a small project (Wainer, Hambleton, & Meara, in progress) to extend the earlier work of Hambleton and Slater (1995, in press) and Wainer (1996, 1997a, 1997b). This new study seeks to take a diverse mix of current NAEP data displays that seem problematic, revise them along the lines of emerging data-reporting principles, and field-test both the original and revised displays with educators and policymakers. The data displays for the study were selected from the 1994 NAEP Reading Study (Williams, et al., 1995). An example of an original display (without color) appears in figure 5.5, and a revised display designed by Howard Wainer appears in figure 5.6. (Four additional original displays and their revisions are also used in the study.)

Figure 5.6. Average reading proficiency by grade and by region—NAEP 1992 and 1994



Source: Wainer, Hambleton, & Meara, in progress

The original and revised displays shown in figures 5.5 and 5.6 will be distributed to policymakers and educators in an interview format, along with the following types of questions:

1. What was the general direction of results between 1992 and 1994?
2. Which region showed the greatest decline in performance for 12th graders from 1992 to 1994?
3. In 1994, which region of the country had the lowest average reading proficiency at all three grade levels?
4. What is the ranking of the regions from *best to worst* in terms of *average reading proficiency* for grade 12 in 1994?
5. Which of the four regions is most typical of the U.S. results?

6. Which regions for 8th graders in 1994 performed better than the average for all of the United States?
7. In everyday language, what do you think is meant by the phrase "significant decrease?"

The answers given by educators and policymakers to these seven questions, using the two versions of the data displays, will be central to the study. In addition, the time needed to respond will be a dependent variable for some questions.

Keep presentation clear, simple, and uncluttered

In the Wainer, Hambleton, and Meara study, five displays from the 1994 NAEP Reading Assessment Study were revised. Participants in the study will be randomly assigned to answer a set of questions about each display using one of the two versions for each display. The findings from the study should determine whether the presentations of NAEP data were improved. Results from the study will be available in summer 1998.

Hambleton, Slater, Allalouf Instructional Module on Score Reporting

Recognizing the need for steps and guidelines in the preparation of data displays and as a follow-up to the work of Hambleton and Slater (1995, in press), Hambleton, Slater, and Allalouf (in progress) are producing an instructional module with a five-step model that follows specific guidelines for preparing tables and figures. An outline of their guidelines for preparing data displays follows:

1. Keep presentation clear, simple, and uncluttered.
    - — Frame the graph on all four sides.
    - — Use no more tick marks than necessary and place ticks on all four sides of the graph frame.
    - — Do not automatically place tick marks or numbers at the corners of the graph frame.
    - — Clearly label the left and bottom axes.
    - — Avoid the use of scale breaks.
    - — Use visually distinguishable symbols. Avoid placing isolated data points too close to the frame where they may be hidden. Identify overlapping data points.
    - — Consider using visual summaries, such as regression lines or smoothing, if the amount of data is large or if important trends are clouded by clutter or unduly influenced by outliers.
    - — Use reference marks to highlight an important value.
    - — Label the elements of bars and graphs (do not use keys or grids). Horizontal bars give room for labels and figures on or near the elements, which is why they are preferred to vertical bars.
    - — Use a table with round numbers in numerical form (5,000, not 5,422 or 5 thousand) if memory of specific amounts is required.
2. Ensure that the graph can stand alone (i.e., a graph should be able to be interpreted in isolation from the main text).
    - — Highlight data, not extraneous material.
    - — Minimize noninformative material in the data region.

   - Ensure that the entire graph can be reduced and reproduced without loss of clarity.

3. Ensure that text complements and supports the graph.
   - Never present numerical data in text form if more than one or two items are to be presented.
   - Use questions following the table or chart to emphasize its chief features.
   - Include text to support and improve interpretation of charts and tables.
4. Plan the graphical presentation.
   - Field-test a sample graph with the intended audience before producing the completed version.
   - Consider using more than one graph to communicate an idea or concept.
5. No form of graph is more effective in all respects than all other forms. However, the following suggestions have been found in the literature:
   - Comparisons based on bar charts are more accurate than comparisons based on circles or squares.
   - Comparisons based on circles or squares are more accurate than comparisons based on cubes.
   - Bar charts prove easier to read than line graphs.
   - Grouped line graphs (each element originating with baseline) are easier to read than segmented line graphs, which prove very difficult.
   - Line codes for graphs should be chosen to minimize confusion.
   - For reading points, multiple lines and multiple graphs are equally good. For comparison, the multiple-line display is always superior.
   - In general, color coding improves performance over the black-and-white code, especially for multiple-line graphs.

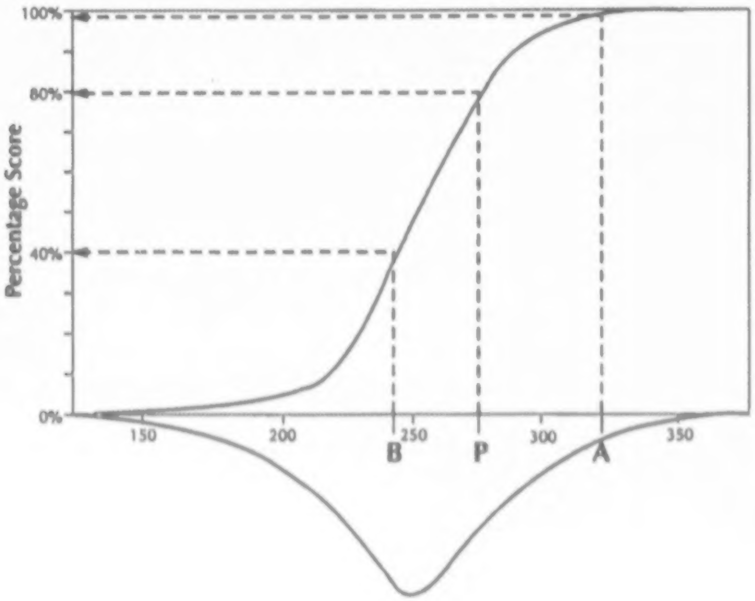
This instructional module, which is scheduled to appear in *Educational Measurement: Issues and Practice* in 1998, and which will include the above guidelines with explanations and examples, is expected to prove valuable to district, state, and national agencies that design data displays and communicate test results.

Market-Basket Reporting

The price of a market basket of food is reported … a single, easy-to-understand measure of economic change.

Mislevy, Bock, and Thissen have created the concept of a "market basket" for score reporting. Mislevy et al., (1996) provide an excellent explanation of market-basket reporting and some of its variations (see also Mislevy, 1998). Their idea apparently originated in market-basket reporting to explain economic changes over time as reflected by the consumer price index. The price of a market basket of food (with known and fixed grocery items) is reported each month to provide the public with a single, easy-to-understand measure of economic change. The extension to education is as follows: Imagine a collection of test items and performance tasks that measure important educational outcomes. The collection of assessment material would reflect diverse item formats, difficulty levels, cognitive levels within a subject area, and any other dimensions of interest. The quality of education might be monitored by reporting the performance of a national

Figure 5.7. Market-basket assessment



sample of students on the "market basket" of items each year. Many policymakers seem to want a single, clear index regarding the quality of education, much like the consumer price index.

The market-basket items would be clearly explained to the public to enhance the meaning of statements such as:

> In 1996, the average American fourth-grade student obtained 37 out of 50 points on the assessment. This is three points higher than the results reported in 1995.

Alternatively or in addition, standards could be used in reporting:

> An advanced student would need to score 45 points on the assessment. Approximately, 5% of the students in the national sample were judged as advanced. In 1995, only 3% of the students met the standard for being advanced. Thus, evidence of improvement is seen.

Figure 5.7 shows the way NAEP achievement levels or performance standards (B-Basic, P-Proficient, and A-Advanced) might be mapped on the percentage score scale (or test score scale) associated with the market-basket assessment. The monotonically increasing curve involved in the mapping is the TCC for the assessment material in the market basket (Hambleton, Swaminathan, & Rogers, 1991). The TCC links the NAEP proficiency scale and the achievement levels to a more meaningful percentage score scale for the particular items and tasks (i.e., the assessment material) in the market-basket assessment. The NAEP score distribution for the student population of interest can be mapped on to the more meaningful percentage score scale for many NAEP audiences along with the achievement levels and can then be used in score reporting. Although many problems must be overcome, the basic concept of market-basket reporting appears to be attractive to NAEP audiences.

A problem may occur if the market-basket items and tasks are administered or released to the public in reports (as seems desirable to communicate fully the meaning of the results). These items and tasks would be compromised and could not be used in future assessments. Students might perform better on them in the future, not because the quality of education improved, but because the assessment material had become known and was taught to them. For the market-basket concept to work then, an equivalent set of items and tasks must be found for each administration. However, the construction of strictly equivalent forms of a test is a very difficult task, and even minor errors would distort the interpretation of the results. Perhaps only part of the market basket of assessment material could be released after each administration. Which parts and the size of the market basket would need to be determined so that only the suitable amount would be released to the public.

Another problem is that released items and tasks might have unintended effects on curricula and assessments such as a narrowing of the curricula to match the released assessment material and elimination of assessment formats that were not used in the market-basket items and tasks. These and other problems have reasonable solutions, but research will be needed to address them before this concept is ready for implementation.

Conclusions

Considerable evidence is found in the measurement literature to suggest that NAEP scales and score reports are not fully understood by intended audiences. At the same time, many signs indicate that the problems associated with these misunderstandings are being addressed. A review of NAEP reports from 1990 to 1996 shows significant improvements in the clarity of displays. The use of anchor points, achievement levels, benchmarking, and market-basket displays, which were addressed in this paper, appears to be valuable for improving NAEP displays. Ongoing research of these and other innovations will further enhance NAEP reports. The emerging guidelines for data displays from Hambleton, Slater, and Allalouf (in progress) address a pressing need and should improve data displays. At the same time, a considerable amount of research needs to be conducted if NAEP displays are to achieve their purposes.

NAEP reports, in principle, provide policymakers, educators, education writers, and the public with valuable information. However, the reporting agency needs to ensure that reporting scales are meaningful to the intended audiences and that displays are clear and understandable. These efforts will almost certainly require the adoption and implementation of a set of guidelines for reporting, which would include the field-testing of all reports to ensure that they can be interpreted fully and correctly. Special attention, too, must be given to the use of figures and tables, which can convey substantial amounts of data clearly when they are properly designed.

NAEP reports, in principle, provide policymakers, educators, education writers, and the public with valuable information.

References

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17* (2), 191–204.

Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement, 29* (2), 163–176.

Bourque, M. L., Champagne, A. B., & Crissman, S. (1997). *1996 science performance standards: Achievement results for the nation and the states.* Washington, DC: National Assessment Governing Board.

Cleveland, W. S. (1985). *The elements of graphing data.* Monterey, CA: Wadsworth.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice, 10* (3), 3–9, 16.

Hambleton, R. K. (1994). *Scales, scores, and reporting forms to enhance the utility of educational testing.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.

Hambleton, R. K., & Bourque, M. L. (1991). *The levels of mathematics achievement: Initial performance standards for the 1990 NAEP Mathematics Assessment.* Washington, DC: National Assessment Governing Board.

Hambleton, R. K., & Slater, S. (1995). Using performance standards to report national and state assessment data: Are the reports understandable and how can they be improved? *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments,* Volume II (pp. 325–343). Washington, DC: National Assessment Governing Board, National Center for Education Statistics.

Hambleton, R. K., & Slater, S. (in press). Are NAEP executive summary reports understandable to policymakers and educators? *Educational Assessment.*

Hambleton, R. K., Slater, S., & Allalouf, A. (in progress). Steps for preparing displays of educational data. *Educational Measurement: Issues and Practice.*

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Henry, G. T. (1995). *Graphing data: Techniques for display and analysis.* Thousand Oaks, CA: Sage Publications.

Jaeger, R. (1992). General issues in reporting of the NAEP trial state assessment results. In R. Glaser & R. Linn (Eds.), *Assessing student achievement in the states.* Stanford, CA: National Academy of Education (107–109).

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29*, 95–110.

Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey.* Washington, DC: U.S. Government Printing Office.

Koretz, D., & Deibert, E. (1993). *Interpretations of National Assessment of Educational Progress (NAEP) anchor points and achievement levels by the print media in 1991.* Santa Monica, CA: RAND.

Linn, R. L., & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29* (2), 177–194.

Mislevy, R. J. (1998). Implications of market-basket reporting for achievement-level setting. *Applied Measurement in Education, 11* (1), 49–63.

Mislevy, R. J., Forsyth, R., Hambleton, R. K., Linn, R., & Yen, W. (1996). *Design/feasibility team report to the National Assessment Governing Board.* Washington, DC: National Assessment Governing Board.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17,* 131–154.

Mullis, I. V. S. (1991). *The NAEP scale anchoring process for the 1990 mathematics assessment.* Paper presented at the meeting of American Educational Research Association, Chicago.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1993). *Executive summary of the NAEP 1992 mathematics report card for the nation and the states.* Washington, DC: U.S. Department of Education.

Phillips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., & Barton, P. E. (1993). *Interpreting NAEP scales.* Washington, DC: U.S. Department of Education.

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher, 21* (1), 14–23.

Wainer, H. (1996). Using trilinear plots for NAEP data. *Journal of Educational Measurement, 33,* 41–55.

Wainer, H. (1997a). Improving tabular displays: With NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Sciences, 21,* 1–32.

Wainer, H. (1997b). Some multivariate displays for NAEP results. *Psychological Methods, 2* (1), 34–63.

Wainer, H. (1997c). *Visual revelations.* New York: Springer-Verlag.

Wainer, H., Hambleton, R. K., & Meara, K. (in progress). *A comparative study of original and revised displays of NAEP data* (tentative title). A project funded by the National Center for Education Statistics.

Wainer, H., & Thissen, D. (1981). Graphical data analysis. *Annual Review of Psychology, 32,* 191–241.

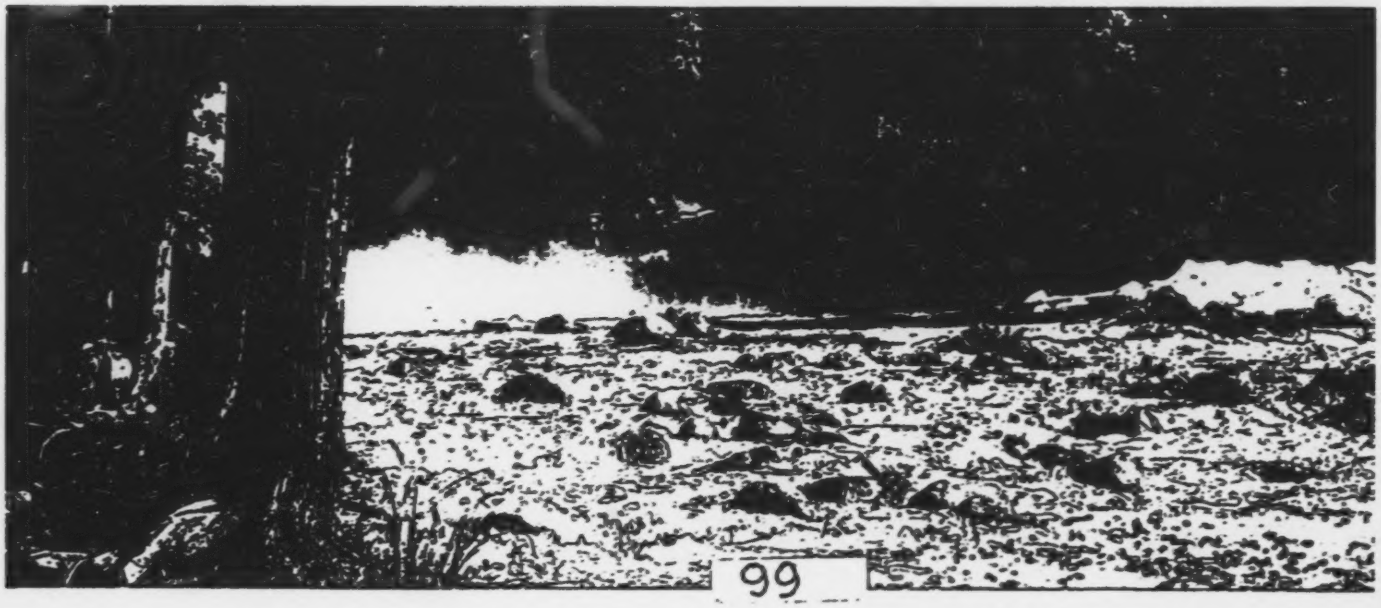
Williams, P. L., Reese, C. M., Campbell, J. R., Mazzeo, J., & Phillips, G. W. (1995). *NAEP 1994 reading: A first look.* Washington, DC: U.S. Department of Education.

SECTION 6

1998 Civics and Writing Level-Setting Methodologies

ACT, Inc. Iowa City, IA

August 1997



1998 Civics and Writing Level-Setting Methodologies

Setting appropriate achievement levels on the National Assessment of Educational Progress (NAEP) helps define some of the important outcomes of education, stating clearly what students should know and be able to do at key grade levels in school. This will make the assessment far more useful to parents and policymakers as a measure of performance in American schools and perhaps as an inducement to higher achievement. The achievement levels will be used for reporting NAEP results in a way that greatly increases their value to the American public.

NAGB 1990

Achievement levels are an important and increasingly integral part of the National Assessment of Educational Progress (NAEP). Achievement levels directly address the way NAEP communicates information about student performance in selected learning areas to a variety of constituencies to improve education in the United States and to meet goals that signal educational parity, at a minimum, in the international arena. In particular, achievement levels give a readily understood means of describing what students should know and be able to do.

The current debate regarding national tests increasingly draws attention to the National Assessment Governing Board (NAGB) and achievement levels in general. Although neither writing nor civics has been proposed for the national tests, an increased interest in the achievement levels set for all NAEP subjects seems inevitable. Thus, the design of this achievement levels-setting (ALS) process is particularly important.

In 1990, NAGB unanimously adopted three achievement levels to serve as the primary means of reporting results for NAEP. The three levels are:

- ***Proficient:*** This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject matter knowledge, applications of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
- ***Advanced:*** This level signifies superior performance beyond *Proficient*.
- ***Basic:*** This level denotes partial mastery of prerequisite knowledge and skills fundamental for *Proficient* work at each grade level.

. . . achievement levels give a readily understood means of describing what students should know and be able to do.

The plan described here is an extension of earlier work by NAGB and ACT, Inc., to set achievement levels in mathematics, reading, writing, geography, U.S. history, and science. ACT's experiences in carrying out the responsibilities of the contract for setting achievement levels on NAEP in these subjects have guided the development of the design features for the 1998 achievement levels-setting process for civics and writing described in this document.

Such questions as what constitutes "civics," "writing," or any other subject assessed by NAEP; how shall the subject be assessed; what information should be collected; and what item development and test administration issues are important, but these are "givens" to this project. Inputs to this process are the following:

- Frameworks for the civics and writing NAEP were developed under contract to NAGB, through a consensus process involving panels of experts and public comment forums held throughout the United States.

Scores are converted to the NAEP scale using a three-parameter item response theory (IRT) model.

- Policy definitions of the achievement levels (given above) were set by NAGB. Preliminary achievement level definitions were developed by framework panelists for each grade level within each subject area—e.g., fourth-grade *Basic* civics, fourth-grade *Proficient* civics, and fourth-grade *Advanced* civics.
- Item development, development of test forms (booklets), and field-testing for each subject area were carried out by the Educational Testing Service (ETS) under contract to the National Center for Education Statistics (NCES). Members of NAGB and the framework consensus panels, among others, participated in the selection of items to be included on the assessment of each subject and at each grade level.
- Assessments of students throughout the nation were selected through samples of schools. The assessments in civics and writing will be administered during the first 3 months of 1998. NCES has authority to administer NAEP. ETS develops the assessment, and Westat administers the assessment under contract to NCES.
- Assessments are scored by National Computer Systems (NCS) under contract to ETS. Scores are reported to ETS where they are converted to the NAEP scale using a three-parameter item response theory (IRT) model. The item parameters and other data, such as scale transformations and student performance data, are provided to ACT for use in setting achievement levels in each assessment subject area.

An ALS process will be conducted to set achievement levels in both civics and writing to be assessed in 1998. This process is designed to produce three products: descriptions of the knowledge and skills that students in each grade level assessed in NAEP (4th, 8th, and 12th) should have to be classified as performing at each level of achievement; the numerical score (or "cutpoint") associated with that level of performance on the particular assessment administered; and test items with (written) responses illustrative of the kinds of knowledge and skills required of students performing at each level of achievement. Outcomes of this process will be:

- Content-based descriptions of each level of achievement for each of the three grade levels assessed by NAEP.
- Numerical cutscores (the lower bound at each achievement level)

that tie these descriptions to performance on the assessments.

- Items available from the 1998 pool of items slated for public release, illustrative of the skills and knowledge characterizing student performance at each achievement level at each of the three grade levels assessed by NAEP in these subjects.

The NAEP score associated with the cutpoints and performance at or above each level tend to be the focal point of NAEP reporting. In addition to deriving recommended cutpoints for *Basic, Proficient,* and *Advanced* achievement levels, important outcomes of this project are the descriptions of the proposed achievement levels and the sample items and responses selected to represent student performance at each achievement level.

The frameworks for the 1998 NAEP assessments include preliminary definitions of the three achievement levels for each grade. As a part of the framework, preliminary definitions were used to guide the development of the assessment items and tasks. ACT has designed a process by which the achievement levels descriptions will be reviewed and finalized prior to convening the ALS panels.

A series of focus groups, one in each of the four NAEP regions, has been planned. These focus groups are being held to collect recommendations for modifications in the preliminary achievement levels descriptions that would increase the *reasonableness* of the descriptions with respect to the policy definitions. The plan is to convene a panel of experts to review the recommendations and make changes deemed appropriate, with respect to the framework. These finalized descriptions will then be presented to NAGB for approval before the ALS pilot studies are conducted.

Because relatively few blocks of items in each assessment will be released for public review, the maximum feasible number of items will be selected for consideration as items to represent student performance at each achievement level. These achievement level descriptions and the illustrative items for each will play prominent roles in communicating the achievement of students on NAEP.

ACT intends to elicit and engage participation by numerous experts, interested organizations, and individuals. The final product will benefit from the input of these individuals and interests. ACT will provide the impetus for the accumulation of information and expertise to be focused on and channeled into the development of the achievement levels and the validation of their interpretations of student performance.

The achievement levels will be developed by a group of individuals representative of both the educational community and the general public.

The achievement levels will be developed by a group of individuals representative of both the educational community and the general public. A broadly representative set of panelists will be identified for each content area. The involvement and participation of stakeholder groups and other interested constituencies will be elicited in all phases of this project. Further, the recommended achievement levels—descriptions, numerical values, and illustrative items—will be made available for public review, and ACT will attempt to engage the public in this review. The primary mode of collecting opinions from these persons and organizations will be through the Internet and Web sites for both ACT and NAGB. All comments offered during this public review phase will be shared with the Technical Advisory Committee on Standards Setting (TACSS) and NAGB, and these comments will become a part of the information compiled to develop the recommendations for NAGB regarding achievement levels in each content area.

Key Points in ACT's Design

ACT has extensive experience in assisting national organizations in determining criterion scores or standards for their programs. ACT has knowledge of current advances in ALS processes and the creativity and expertise to modify and expand upon these as appropriate for specific implementations. In addition, ACT designed and implemented the achievement levels-setting process for the 1992 NAEP in mathematics, writing, and reading, the 1994 NAEP in geography and U.S. history, and the 1996 NAEP in science. We believe that this experience provides the insights and understanding, merged with the technical and methodological expertise and experience, for designing a process to fully address the many challenges and requirements of the NAEP assessments.

The ALS process, first implemented in 1992, incorporated state-of-the-art methodologies in standard setting and fully accomplished the goals required of a successful standard-setting process.

In designing these processes, ACT has carefully reviewed the procedures previously implemented in setting achievement levels for NAEP and those described in recent literature on standard setting. We have attempted to evaluate objectively the primary features of the processes implemented for NAEP and to identify ways to improve upon them. We are aware of the features that are unique to NAEP. Many NAEP features require special consideration in the ALS design, and some standard-setting procedures simply will not work for NAEP.

Many features of earlier processes have been retained; many others have been improved upon and enhanced. A sampling plan for identifying and recruiting panelists was successfully designed and implemented for the 1992, 1994, and 1996 ALS processes. The sampling plan has yielded broadly representative panels of well-qualified individuals. The approval of a diverse set of interested individuals, organizations, and groups was sought and achieved for both the sampling plan and the overall research design.

The ALS process, first implemented in 1992, incorporated state-of-the-art methodologies in standard setting and fully accomplished the goals required of a successful standard-setting process. The process designed by ACT for the 1994 and 1996 NAEP ALS process incorporates all features recognized as "desirable" during the conference on standard setting sponsored by NAGB and NCES in 1994 except sharing consequences data with panelists during the ALS process. The current design calls for a change in NAGB policy to allow this addition.

Questions and concerns have emerged regarding psychometric and standard-setting issues related to NAEP achievement levels that have never been addressed. ACT has raised several of these, and we have attempted to openly address all that were brought to our attention. The design presented in this document incorporates improvements and enhancements generated through experiences gained during previous achievement levels-setting efforts for NAEP.

Key features of the 1998 proposal include the following:

1. A *sampling plan* for recruiting panelists for each ALS meeting that will result in the involvement of a well-qualified, representative panel of judges, while introducing efficiencies resulting from the experiences of identifying and recruiting panelists for many previous NAEP panels.
2. A series of *focus groups* to make recommendations that will be incorporated into the preliminary achievement levels descriptions. These modified descriptions will be presented to NAGB for final approval prior to convening ALS panelists.
3. A *research agenda* incorporated into the general approach of the ALS process and validation process that will contribute significantly to the product of the 1998 achievement levels-setting process *and* to the body of knowledge in standard setting, item response theory, and other technical and methodological areas of educational assessment and measurement.
4. Field trials (two separate studies each) for both civics and writing to test rating methodologies and other features of the proposed design and to collect research data regarding the design *before* the pilot studies are implemented.
5. *Pilot studies* for both civics and writing that will provide the opportunity for testing the process and making needed changes and adjustments prior to implementing the ALS process.
6. *Ample time* in the agenda for the ALS pilot studies and meetings to successfully address key elements of the ALS process.
7. Extensive *training for panelists,* not only in the methods of evaluating items and rating them to set achievement levels but also in the consequences of those standards, in the objectives and purposes of NAEP and NAGB, and in educational assessment issues and policies.

> *Pilot studies . . . will provide the opportunity for testing the process and making needed changes and adjustments. . . .*

8. A more *deliberative process* engaging panelists in two different methods for arriving at the numerical cutscores.
9. Consequences data provided to panelists *during* the process so that cutscores may be adjusted after panelists have been informed about the consequences of setting achievement levels at specific score points.
10. *Customized computer software* that uses the IRT calibrations of the NAEP items and scale to produce on-site feedback to panelists on the consistency and convergence of ratings and on the consequences of their ratings.

11. *Scannable rating forms* to reduce the time required to enter and analyze data and produce feedback for panelists to use during the process.
12. The *same, well-trained process facilitation staff* for each pilot study and each ALS meeting to ensure consistency in implementation.
13. *Content facilitation staff* well versed in the NAEP framework or item pool for the subject (or both) to work with each grade-level panel on the pilot studies and the ALS in each subject.
14. *On-site logistic planning and support services* using full-time, ALS-experienced project staff.
15. A team of veterans represented on the project staff, the internal advisory team, and the external committee of technical advisors, including *highly experienced experts* in standard-setting methodology, psychometrics, sampling statistics, writing and other educational assessment, collective decision making, and meeting management, joined by new members with fresh insights and ideas to enrich the outcomes of the project.

ACT's general approach to deriving recommended achievement levels for the NAEP is guided by five overarching principles:

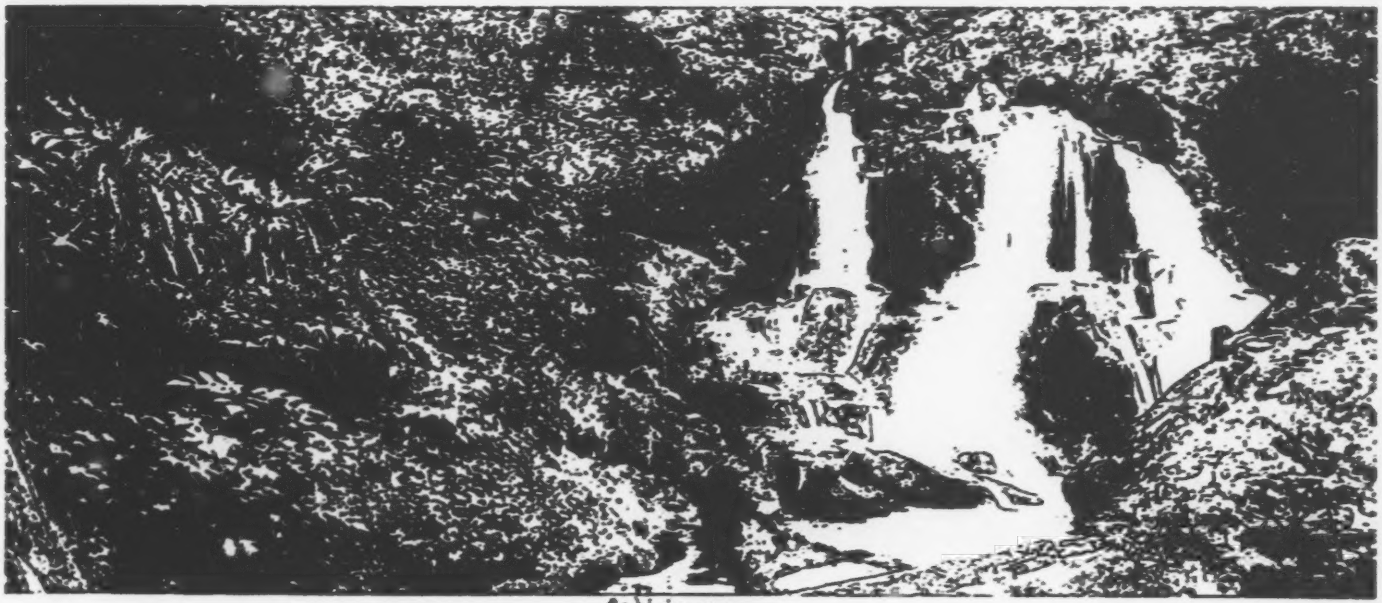
1. There must be broad, thorough, and open participation from all relevant populations in the ALS process.
2. Highly sensitive and confidential materials, reports, and information must be handled in an appropriate manner.
3. The levels-setting process must be carefully designed, technically sound, rigorously implemented, and appropriately validated.
4. The levels-setting process must be comprehensible to interested parties and easily implemented by process participants.
5. NAGB must exercise informed direction over all major project activities and be kept fully apprised of all relevant project information.

SECTION 7

The Criticality of Consequences in Standard Setting: Six Lessons Learned the Hard Way by a Standard-Setting Abettor

W. James Popham University of California, Los Angeles

August 1997



The Criticality of Consequences in Standard Setting: Six Lessons Learned the Hard Way by a Standard-Setting Abettor

When I was initially invited to present some ideas to members of the National Assessment Governing Board (NAGB) in August 1997, my topic was to be "consequential validity." I had written about consequential validity a few months earlier, contending that I did not regard it as a particularly commendable concept. However, as I later learned that the August NAGB session was to focus directly on that group's standard-setting activities, I decided to emphasize the role of consequences in the setting of performance standards. Although I regard consequential validity as a psychometrically sordid idea, the impact of consequences on standard setting is, and *should be,* enormous.

I decided to draw on my experiences in the setting of standards for more than three dozen high-stakes tests for students, teachers, and administrators. I would like to describe some lessons that I have learned.

What to Call Oneself?

On careful consideration, I realized that I had never set one of these performance standards. Most typically, I served as moderator for a statewide standard-setting panel. In other settings, I functioned as an advisor to the ultimate standard setters. Clearly, I had to find a suitable descriptor for my role in the aforementioned standard-setting endeavors. I toyed with such possibilities as *consultant, advisor,* and *coconspirator,* but none seemed on the mark. But then I looked up the meaning of *abettor* in *Webster's Collegiate Dictionary*; an abettor's role is "to encourage, support, or countenance by aid or approval, usually in *wrongdoing*" [italics added]. Obviously, I had found an appropriate label.

Six Lessons

Six consequence-relevant lessons that this standard-setting abettor has learned while wrestling with the establishment of performance standards are found below. Each lesson is followed by a brief comment.

- Lesson 1. The chief determiner of performance standards is not truth; it is consequences.

Lesson 1. The chief determiner of performance standards is not truth; it is consequences.

Abettor's Comment: If an explanation is sufficiently important to warrant a formal standard-setting effort, it is invariably true that meaningful consequences will be linked to examinees' performances. Accordingly, when standard-setting panelists determine performance levels for such examinations, those standard setters are typically influenced more by the consequences of the standards they set (e.g., the number of students who will not receive a high school diploma) than by any notions about true ("correct") performance levels.

- Lesson 2. Any quest for "accurate" performance standards is silly.

Abettor's Comment: There are always multiple consequences linked to performances on significant tests. To illustrate, for a high school graduation test, such consequences would include diploma denials, citizens' estimates of educators' effectiveness, and the business community's satisfaction with diploma recipients. Because different standard setters' perceptions of the significance of those consequences will vary, the performance levels ultimately selected typically reflect a judgmental amalgam rather than a correct performance standard.

Lesson 2: Any quest for "accurate" performance standards is silly.

Unfortunately, some psychometricians have spent so much time searching for true scores and accounting for error variance that they sometimes attempt to impose a truth-and-error paradigm on the standard-setting process. A preferable approach to standard setting would be to recognize that it is fundamentally a consider-the-consequences enterprise.

- Lesson 3. Early in the standard-setting process, all likely consequences should be explicated for standard setters.

Abettor's Comment: Standard setters frequently become so preoccupied with the most obvious and advertised consequence of using a test (e.g., for professional licensure) that they fail to recognize the certainty or possibility of other potentially significant consequences. If standard setters are alerted to the full range of likely consequences of a test's use, they will be more apt to function thoughtfully by considering all consequences, or at least those consequences they consider important.

- Lesson 4. If certain standard setters (because of their positions or affiliations) are apt to be biased in their judgments, such potential biases should be identified early in the standard-setting process.

Abettor's Comment: In many attempts to establish performance standards, particular standard setters enter the process frequently with powerful biases in favor of higher or lower performance standards. Such biases frequently flow from the orientation of the entity (e.g., teachers union) represented by a standard setter. Often, these blatantly biased individuals will profess nonpartisanship during standard-setting deliberations.

Those directing the standard-setting enterprise should isolate such biases at the outset of the deliberations. This could be done by identifying the quite normal proclivity of certain categories of standard setters (no names) to favor performance standards compatible with preferences of the groups those standard setters represent. Visible biases are more readily countered than are camouflaged biases.

- Lesson 5. When it is hoped that a testing program will stimulate subsequent increased proficiency among those to be tested, incremental elevations of performance levels, over a period of time, can avoid undesirable consequences.

Abettor's Comment: Frequently, the skills or knowledge assessed by a new test will reflect higher-level proficiencies that it is hoped will be possessed by future examinees. The examination system is being used to spur the acquisition of these more demanding capabilities. Yet, because few examinees will, in the new testing program's early days, possess the ultimately desired proficiency levels, standard setters are sometimes tempted to opt for lower performance standards to avoid penalizing these early test takers. If such low performance standards are allowed to persist throughout the duration of the testing program, however, its lofty improvement aspirations will not be realized. This classic approach-avoidance conflict can often be circumvented through the use of preannounced, incremental increases in required performance levels over time.

- Lesson 6. Performance-level descriptors must accurately communicate, in an intuitively comprehensible fashion, to all concerned constituencies.

Abettor's Comment: Those individuals most actively involved in the development and operation of testing programs, or in the determination of performance standards, often become so familiar with the nuances of what is being assessed that they devise sophisticated performance-level descriptive schemes not readily understandable to the uninitiated. Sometimes, for example, exotic scale-score reporting systems are created with the thinly veiled purpose of obfuscating what would be regarded by the public as unacceptably low performance standards. Performance standards must be readily understandable to those who are concerned about examinees' performances.

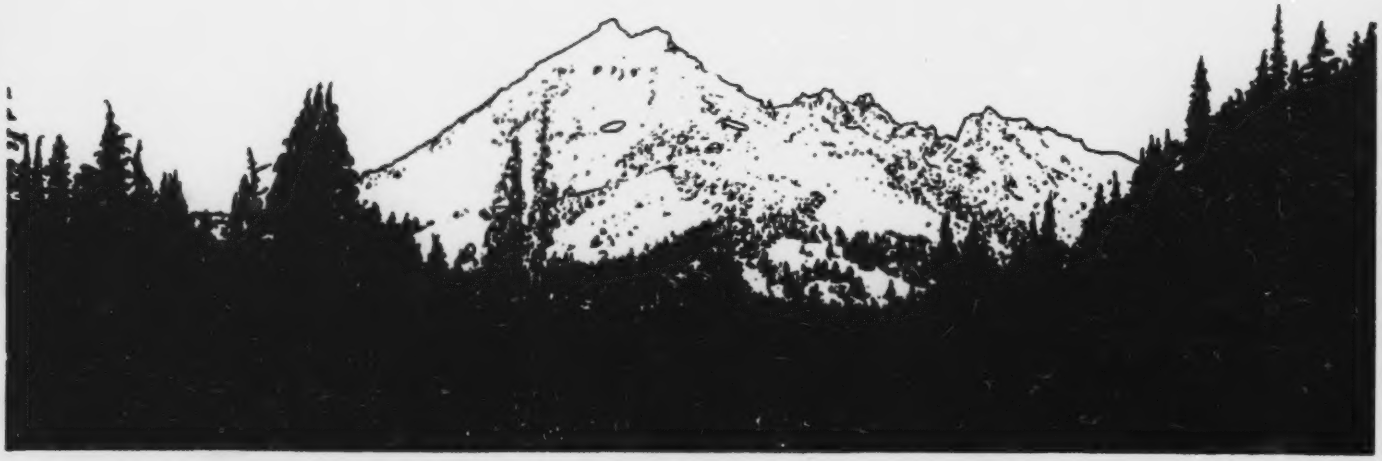
Lesson 6. Performance-level descriptors must accurately communicate, in an intuitively comprehensible fashion, to all concerned constituencies.

A Final Admonition

Many important educational decisions are made without a careful decision-making process. Standard setting for high-stakes tests, however, should never be made in the absence of thoughtful, systemized judgment. My understanding of the standard-setting procedures previously employed by NAGB is that there has been far too much deference given to the quantitative "truth-seekers." Even though there will never be a standard-setting machine that pumps out unflawed performance levels, all we can ask is that NAGB and other standard setters circumspectly consider the consequences of the performance levels they set.

SECTION 8

Acknowledgements and Appendices



Acknowledgements

This Achievement Levels Workshop involved the support and assistance of numerous staff from the National Assessment Governing Board (NAGB), Educational Testing Service (ETS), ACT, Inc., and Aspen Systems Corporation. The committee would like to acknowledge the assistance of each staff member who helped to arrange and plan for this inaugural event.

In addition, the committee wishes to thank the authors whose work appears in this volume of the proceedings. Their informative presentations and learned papers provided a collection of some of the best and current thinking about achievement levels. We would especially like to thank Robert A. Forsyth, University of Iowa; Wim J. van der Linden, University of Twente, The Netherlands; Ronald K. Hambleton, University of Massachusetts; David Thissen and his colleagues, University of North Carolina; W. James Popham, University of California, Los Angeles/IOX; and Susan Loomis, ACT NAEP Project Director.

We would also like to thank the attendees who offered special comments and insights on the various topics on the agenda, including Lauress Wise, member of the National Academy of Sciences NAEP Evaluation Panel; Jon Cohen from the American Institutes of Research (AIR), a staff member of the Advisory Committee for Evaluation Statistics (ACES); and William Brown, member of the Technical Advisory Committee on Standards Setting (TACSS).

Special thanks goes to Munira Mwalimu and the staff at Aspen for their logistical support and Jewel Bell of the NAGB staff for her assistance in preparing briefing materials, attending to participant onsite needs, and helping the editor in preparing these proceedings.

Appendix A: Participants

Luz Bay
ACT, Inc.
2201 North Dodge Street
Iowa City, IA 52243
Telephone: (319) 337-1639
Fax: (319) 337-3020

Jewel Bell
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6938
Fax: (202) 357-6945

Mary Blanton
Attorney
Blanton & Blanton
228 West Council Street
Salisbury, NC 28145-2327
Telephone: (704) 637-1100
Fax: (704) 637-1500

Mary Lyn Bourque
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6940
Fax: (202) 357-6945

William Brown
University of North Carolina,
Chapel Hill
121 Dunedin Court
Cary, NC 27511
Telephone: (919) 467-2404
Fax: (919) 966-6761

James Carlson
Educational Testing Center
Rosedale Road, Box 6710
Princeton, NJ 08541-6170
Telephone: (609) 734-1427
Fax: (609) 734-5410

Peggy Carr
National Center for Education
Statistics
555 New Jersey Avenue, NW
Washington, DC 20208
Telephone: (202) 219-1761
Fax: (202) 219-1801

Lee Chin
ACT, Inc.
2201 North Dodge Street
Iowa City, IA 52243
Telephone: (319) 337-1639
Fax: (319) 339-3020

Jon Cohen
Chief Statistician
American Institutes for Research
1000 Thomas Jefferson Street, NW
Washington, DC 20007
Telephone: (202) 944-5300
Fax: (202) 344-5408

Mary Crovo
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002-4233
Telephone: (202) 357-6938
Fax: (202) 357-6945

Barbara Dodd
University of Texas
Measurement and Evaluation Center
2616 Wichita Street
Austin, TX 78705
Telephone: (512) 232-2654
Fax: (512) 471-3509

James Ellingson
Fourth-Grade Teacher
Probstfield Elementary School
2410 14th Street, South
Moorhead, MN 56560
Telephone: (218) 299-6252
Fax: (701) 298-1509

Thomas Fisher
Administrator
Department of Education
325 West Gaines Street, FEC 701
Tallahassee, FL 32399–0400
Telephone: (904) 488–8198
Fax: (904) 487–1889

Robert A. Forsyth
Professor
College of Education
University of Iowa
320 Lindquist Center
Iowa City, IA 52242
Telephone: (319) 335–5412
Fax: (319) 335–6038

Michael Guerra
Executive Director
National Catholic Education Association
Secondary School Department
1077 30th Street, NW
Suite 100
Washington, DC 20007
Telephone: (202) 337–6232
Fax: (202) 333–6706

Ronald K. Hambleton
Professor of Education
University of Massachusetts, Amherst
Hills House South–Room 152
Amherst, MA 01003
Telephone: (413) 545–0262
Fax: (413) 545–4181

Patricia Hanick
ACT, Inc.
2201 North Dodge Street
Iowa City, IA 52243
Telephone: (319) 337–1639
Fax: (319) 337–3020

Eugene Johnson
Educational Testing Center
Rosedale Road
Princeton, NJ 08541–6170
Telephone: (609) 734–5598
Fax: (609) 734–5420

Susan Loomis
ACT, Inc.
2201 North Dodge Street
Iowa City, IA 52243
Telephone: (319) 337–1048
Fax: (319) 337–3021

John Mazzeo
Educational Testing Center
Rosedale Road
Princeton, NJ 08541–6170
Telephone: (609) 734–5298
Fax: (609) 734–5410

Karen Mitchell
National Academy of Sciences
2101 Constitution Avenue, NW
Room HA178
Washington, DC 20418
Telephone: (202) 334–3407
Fax: (202) 334–3584

Mark Musick
President
Southern Regional Education Board
592 10th Street, NW
Atlanta, GA 30318–5790
Telephone: (404) 875–9211
Fax: (404) 872–1477

Michael Nettles
Professor of Education and
Public Policy
University of Michigan
2035 School of Education Building
610 East University
Ann Arbor, MI 48109–1259
Telephone: (313) 767–1982
Fax: (313) 764–2510

Christine O'Sullivan
Educational Testing Service
Rosedale Road
Princeton, NJ 08541–6170
Telephone: (609) 734–5598
Fax: (609) 734–5420

Norma Paulus
Superintendent of Public Instruction
State Department of Education
255 Capitol Street, NE
Salem, OR 97310–0203
Telephone: (503) 378–3573, ext. 526
Fax: (503) 378–4772

W. James Popham
Director of IOX Assessment
Associate Professor Emeritus
University of California, Los Angeles
5301 Beethoven Street
Suite 109
Los Angeles, CA 90066
Telephone: (310) 545–8761
Fax: (310) 822–0269

William Randall
Connect
1580 Logan Street
Suite 740
Denver, CO 80203
Telephone: (303) 894–2146
Fax: (303) 894–2141

Douglas Rindone
Connecticut State Department
of Education
P.O. Box 2219
165 Capitol Avenue
Hartford, CT 06145
Telephone: (203) 566–1684
Fax: (203) 566–1625

Andrea Schneider (representative)
Office of Governor Roy Romer
136 State Capitol Building
200 East Colfax Avenue
Denver, CO 80203
Telephone: (303) 866–2471
Fax: (303) 866–2003

Sharif Shakrani
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002–4233
Telephone: (202) 357–6938
Fax: (202) 357–6945

David Thissen
Professor of Psychology
University of North Carolina
CB# 3270 Davie Hall
Chapel Hill, NC 27599
Telephone: (919) 962–5036
Fax: (919) 962–2537

Roy Truby
National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002–4233
Telephone: (202) 357–6938
Fax: (202) 357–6945

Wim J. van der Linden
Professor of Educational
Measurement and Data Analysis
Facility of Educational Science
and Technology
University of Twente
7500 AE Enschede
The Netherlands
Telephone: 011–31–53–893–581
Fax: 011–31–53–48–928895

Deborah Voltz
Assistant Professor
Department of Special Education
University of Louisville
154 Educational Building
Louisville, KY 40292
Telephone: (502) 852–0561
Fax: (502) 852–1419

Lisa Weil (representative)
Office of Governor Roy Romer
136 State Capitol Building
200 East Colfax Avenue
Denver, CO 80203
Telephone: (303) 866-2471
Fax: (303) 866-2003

Paul Williams
Educational Testing Service
Rosedale Road
Princeton, NJ 08541-6170
Telephone: (609) 734-1427
Fax: (609) 734-1878

Lauress Wise
President
HummRo
66 Canal Center Plaza
Suite 400
Alexandria, VA 22314
Telephone: (703) 549-3611
Fax: (703) 549-9025

Rebecca Zwick
University of California, Santa Barbara
Department of Education
Santa Barbara, CA 93106-9490
Telephone: (805) 893-7762
Fax: (805) 893-7264

Appendix B: Conference Agenda

Wednesday, August 20

1:00 p.m.	Welcome and Introductions
1:15 p.m.	Working Lunch
2:15 p.m.	***Frameworks and Specifications: Impact on Achievement Levels*** Robert A. Forsyth University of Iowa
3:45 p.m.	Break
4:00 p.m.	Discussion
5:00 p.m.	Adjourn
6:00 p.m.	Working Dinner, Hotel Gardens

Thursday, August 21

8:00 a.m.	Continental breakfast
8:30 a.m.	***Test Construction/Preliminary Descriptions: Impact on Achievement Levels*** Wim J. van der Linden Law School Admissions Services and University of Twente, The Netherlands
10:00 a.m.	Break
10:15 a.m.	***Judgement Scoring: Impact on Achievement Levels*** Ronald K. Hambleton University of Massachusetts, Amherst
Noon	Lunch, Millennium Room
1:00 p.m.	Discussion
2:00 p.m.	***Mixed Item Formats: Impact on Setting Standards*** David Thissen University of North Carolina, Chapel Hill
3:30 p.m.	Break
3:45 p.m.	***Validity/Reliability Issues for NAEP Achievement Levels*** Jon Cohen American Institutes for Research
4:45 p.m.	Discussion
5:30 p.m.	Adjourn (dinner on your own)

Friday, August 22

7:30 a.m.	Continental Breakfast
8:00 a.m.	***Consequential Validity for Setting Standards*** W. James Popham University of California, Los Angeles/IOX
9:30 a.m.	Discussion
10:00 a.m.	Break
10:15 a.m.	***Civics and Writing Level-Setting Methodologies*** NAEP Staff ACT
Noon	Lunch, Millennium Room
1:00 p.m.	***Wrap-up Panel Discussion*** William Brown Brownstar, Inc. Cary, NC W. James Popham University of California, Los Angeles/IOX Lauress Wise National Academy of Sciences
2:00 p.m.	Adjourn

NATIONAL ASSESSMENT GOVERNING BOARD